Evaluating student success in general education programs using mixed linear models

Clare Cruz

Department of Statistics and Data Science

Carnegie Mellon University

clarecru@andrew.cmu.edu

Abstract

The Dietrich College at Carnegie Mellon University is interested in evaluating the student performance in their new general education program. This study aims to analyze the recent experimentation performed by the college to see the associations and distributions between the ratings. The data for the experiment consists of a sample of 91 student project papers or "artifacts" from freshmen statistics courses from the 2019 calendar year that was rated by three raters using seven different rubrics. The ratings were evaluated using descriptive statistics, intra-cluster correlations, percent of exact agreement, and multiple mixed linear models. In the results, two rubrics received lower scores and one rater tended to give lower scores. Each rater also had one rubric where they had ratings that significantly disagreed with the other raters. Then, the mixed-effects models suggest that rater and semester affect the ratings for certain rubrics and that the raters have different interpretations of the rubrics and artifacts.

Introduction

The Dietrich College at Carnegie Mellon University is administering a new "General Education" program for undergraduate students. In this program, all undergraduates must take a specific set of courses and experiences. To evaluate the success of the program, the college is aspiring to rate student performance in all the "Gen Ed" courses year each. Recently, the college has been experimenting with student evaluation with the freshman statistics courses using raters from across the college. The dean of the college is interested in the results of the experimentation that are outlined in four key research questions:

1.) **Rater and Rubric Rating Distributions – Is** the 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.) **Rater Agreement –** For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?

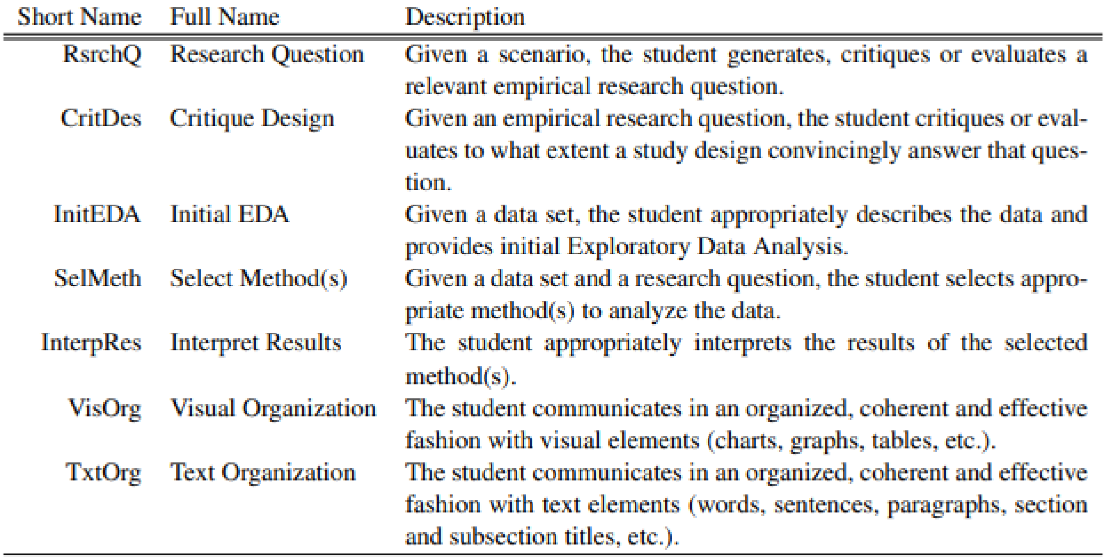
3.) **Rating's Factors –** More generally, how are the various factors in this experiment (Rater, Semester, Sex, Rubric) related to the ratings? Do the factors interact in any interesting ways?

4.) **Unique Rating Factors** – Are there any unique factors that relate to ratings in any of the rubrics? If so, what are some possible explanations for the disparity?

Data

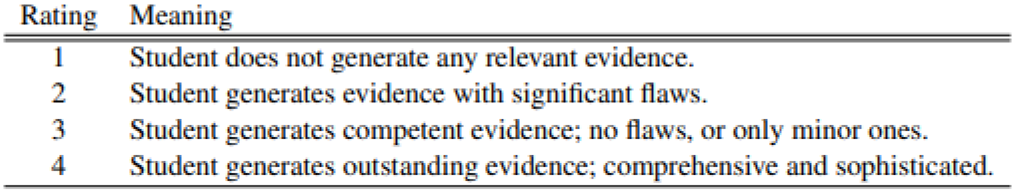
The data set for this study comes from a new experiment performed by Dietrich College. In the experiment, 91 project papers, called "artifacts", were randomly sampled from the fall and spring section of the freshman statistics course for the 2019 calendar year. To evaluate these artifacts, three raters from three different departments were asked to rate the artifacts using seven separate rubrics. The seven rubrics generally follow a particular aspect of a research paper and can be viewed in detail in Table 1. For all the rubrics, the rating scale is the same with values ranging from integers between 1 to 4. The criteria for each rating can be found in Table 2.

Table 1: Descriptions of the seven artifact evaluation rubrics.



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

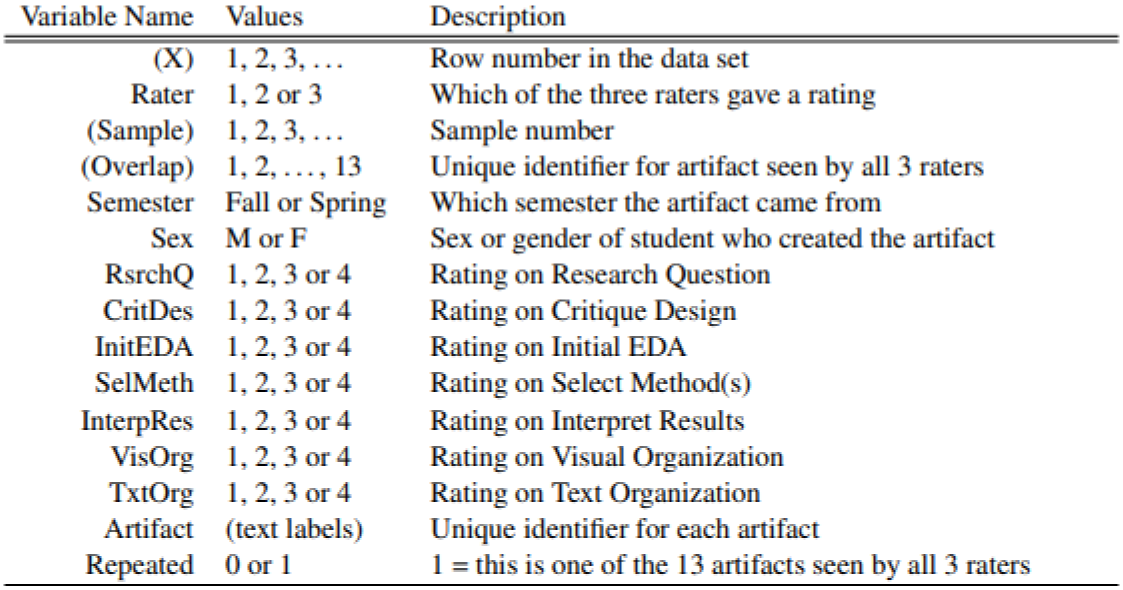
Table 2: Rating scale used for all evaluation rubrics.



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.

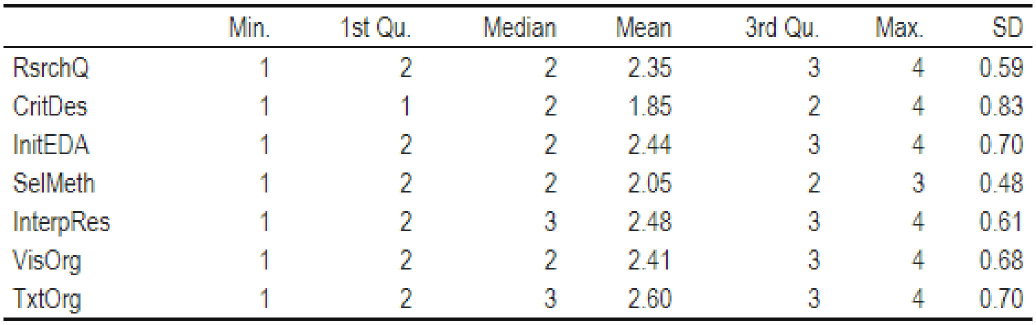
To help evaluate the performance of the raters, thirteen of the artifacts were rated by all three raters. The remaining 78 artifacts were rated by only one rater. In other words, every rater rated 39 artifacts and 26 of those were only rated by that rater. Moreover, the dataset for this analysis contains variables for the rater, ratings, and general student information which can all be viewed in detail in table 3. The dataset has been formatted in two different ways to make the analysis easier. The first format of the data is in the ratings.csv file and is organized the same as the information presented in Table 3. The second format is of the data is in the tall.csv and contains the same information as the first data file. But the ratings are in one column which makes the data longer or taller, hence the filename.

Table 3: Variable definitions for the experiment data from Carnegie Mellon University.



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

Table 4: Summary statistics of the ratings for each rubric using the full dataset.



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.35	3	4	0.59
CritDes	1	1	2	1.85	2	4	0.83
InitEDA	1	2	2	2.44	3	4	0.70
SelMeth	1	2	2	2.05	2	3	0.48
InterpRes	1	2	3	2.48	3	4	0.61
VisOrg	1	2	2	2.41	3	4	0.68
TxtOrg	1	2	3	2.60	3	4	0.70

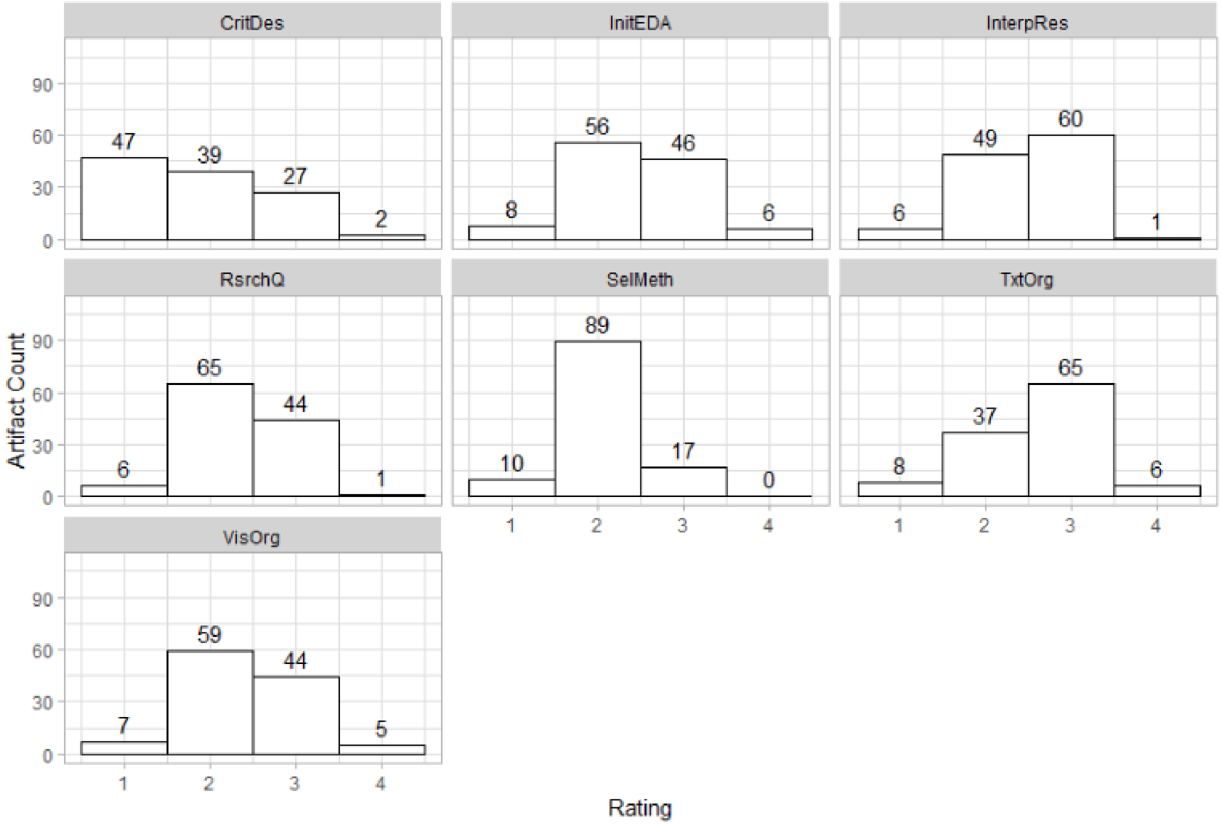
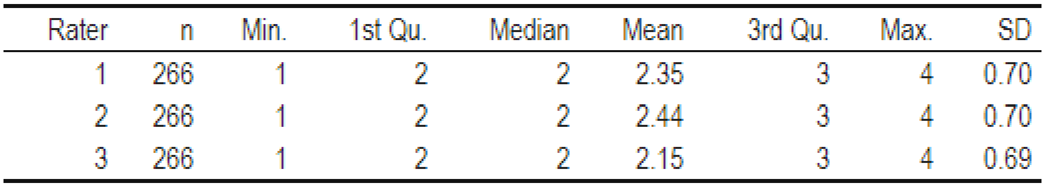


Figure 1: Histograms of all ratings for each rubric using the full dataset.

Each of the four ratings was given to an artifact for every rubric. However, more artifacts were given middle scores such as two or three for most of the rubrics. The critical design rubric is the only rubric where a score of one was given the most. The summary statistics for each rubric using the full dataset are in table 4 and Figure 1. The results for the subset of data are on page 27 in the technical appendix.

Table 5: Summary statistics for the ratings given by each rater using the full dataset.



Rater	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
1	266	1	2	2	2.35	3	4	0.70
2	266	1	2	2	2.44	3	4	0.70
3	266	1	2	2	2.15	3	4	0.69

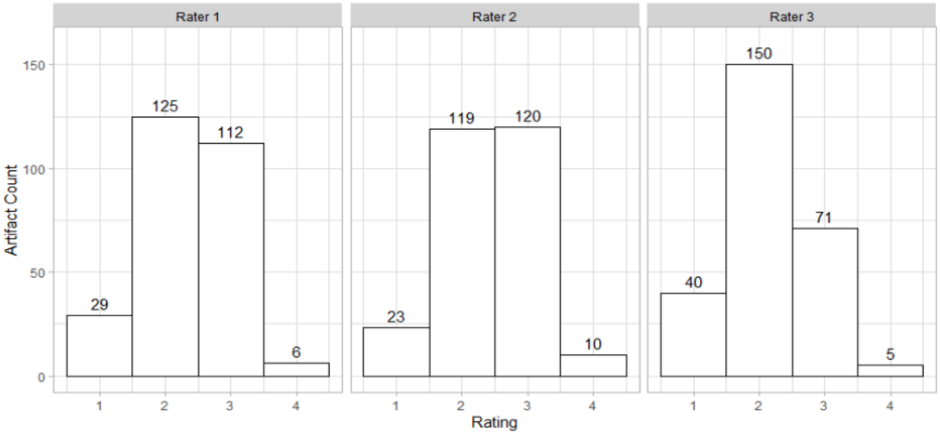


Figure 2: Histograms of all ratings for each rater using the full dataset.

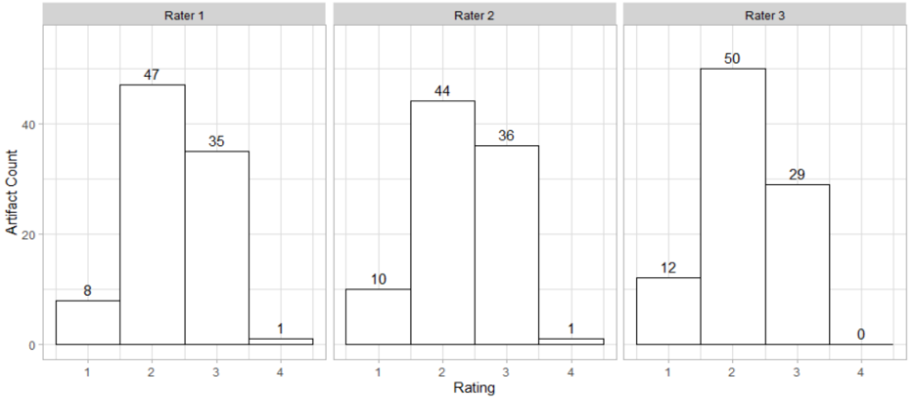


Figure 3: Histograms of all ratings for each rater using the subset of 13 artifacts.

The distribution of the ratings is consistent between the three raters in the subset of data, with most of the raters scoring artifacts as a two or a three (Figure 3). In contrast, the histograms in Figure 2 show that the third rater gave more lower scores than its peers.

Methods

Recall that the dean of Dietrich college is interested in the ratings' associations and distributions in the student evaluation experiment conducted by the college. The methods for each of the dean's questions are outlined here:

1. Rater and Rubric Rating Distribution

The first research question focuses on the distributions of ratings between the seven rubrics and the three raters. In particular, the goal for this section is to if there is any disparity in ratings by rubric or rater. The distributions will be analyzed and compared using descriptive statistics and histograms. The comparison will be performed for the full dataset and the subset of data that contains the 13 artifacts that were rated by all three raters to corroborate conclusions.

2. Rater Agreement

The next goal is to investigate the rating agreement between the three raters to see if there is any disagreement by rubric and which raters are contributing to that disparity. While there are many ways to evaluate agreement, this study will utilize two different methods. The first approach involves calculating the standard intra-cluster correlation (ICC) which measures the average association between values within a certain group. For this analysis, the ICC will be calculated for each rubric using a separate simple mixed model with the artifact as the grouping variable. The methodology will result in seven aggregated correlations for each rubric that represent the association of ratings between the raters. This result is particularly helpful because strong negative or positive correlations indicate that the raters tended to disagree or agree on the ratings. The analysis method will be performed for the subset and full dataset to validate conclusions.

The second method utilizes the percent of exact agreement which determines the proportion of instances where two raters had the same rating for an artifact. Unlike the previous approach, the percent of the exact agreement will provide evidence towards which raters are causing any disagreement. The computation for the percent of the exact agreement involves looking at the two-way tables for the ratings and every pairwise permutation of rater with each rubric, for a total of 21 tables. For each table, the percent of exact agreement is the sum of the main diagonal over the total number of artifacts. This results in 21 percent of exact proportions to be compared. If the percent of exact agreement is high for the three combinations of raters, then there is evidence that the raters agreed on the ratings. But if there is one pair of raters with a lower percentage, then one of those raters is causing the disagreement.

3. Rating's Factors

Another research question asks to investigate the associations between the ratings and the various factors included in the experiment. Again, there are several ways to examine this question but in this analysis two methods will be utilized. The first method involves creating seven different mixed effect models for each rubric with the artifact as the grouping variable. A mixed model is a preferred approach for this question because the experiment contains a grouping or random variable that has shared differences between the groups. Specifically, the artifact is a random grouping variable since the differences in the ratings for any artifact should be similar across artifacts. In other words, if the raters disagree about the ratings for one artifact, then a similar artifact should also have the same disagreement. allows for grouping or random variables and accounts for shared differences between the groups. Thus, creating a mixed effect model for each rubric with the artifact as the grouping variable allows for convenient interpretations of the associations for the ratings.

To start the model-building process, the intercept-only model will be fitted to each of the seven rubric's data. Then, the non-grouping variables or the fixed effects will be added to each of the seven models to see if any of them are significant in predicting the ratings, just like simple linear regression. In particular, the tested fixed effects are sex, semester, and repeated while the ANOVA test with the AIC and BIC will be used to determine the significance of the added effects. The backward elimination selection method will also be used to see which fixed effect should be added to the model as further evidence. Once the important fixed effects have been added to the models, any corresponding interaction terms will be tested for significance using ANOVA and the lmer function. Note that the interaction terms will only be evaluated if the model contains a fixed effect and that the LRT will be excluded from the evaluation when the random effects are assessed. Finally, if any of the seven models contain a fixed effect, the equivalent random effect will also be tested to see if it should be included in the model. The result of the methodology will contain seven mixed effect models which will be evaluated using their summary regression statistics. While these models will be informative of the relationships with ratings, it does not explore the interactions between rubrics since each rubric has it's own model.

Therefore, the other method creates one fixed model so that the interactions terms between the rubric and the other fixed effects can be evaluated. The model building process is the same as the one described for the first method but only one model with all the data is fit. Again, the final model will be analyzed using its summary regression statistics.

4. Unique Rating Factors

Finally, the models from the previous research question will be further examined to see if there are any outlying or unusual associations with the fixed factors. For example, if one of the seven models contains a fixed effect that is not included in the other size models then that model will be investigated to see what could potentially be causing that disparity. The investigation will include statistical summaries and counts based on the model results.

Results

1. Rating and Rubric Rating Distribution

To start, the summary statistics in the full and subset dataset show that the distribution of ratings between rubrics is not the same (Table 4). Specifically, the critical design rubric has lower scores since the 1st and 3rd quartile are both lower than most of the other rubrics and the mean is lowest among the other rubrics. Additionally, the selection method rubric seems to give similar scores since a two was given for at least 50% of the artifacts and this rubric had the smallest standard deviation. Plus, the selection method rubric is the only rubric that did not receive a score of 4.

The histograms of the ratings for each rubric show similar patterns to the summary statistics for the full and subset of data. First, the critical design ratings are right skewed with most artifacts receiving a 1, a trait that is unique to this rubric. Also, the histogram for the selection method rubric shows how an overwhelming majority of artifacts received a 2. The remaining rubrics all show similar peaks for scores 2 and 3. The histograms for the full dataset are in Figure 1 while the figures for the subset of data are on page 27 of the technical appendix. Based on these observations, there is evidence that the ratings for each rubric do not have the same distribution and that the critical design and selection method rubrics tend to have lower scores.

Next, the summary statistics are identical for the raters in the full dataset and the subset of 13 artifacts. The only small deviations are the lower average rating for rater three and rater three did not give a rating of four in the subset. The summary statistics for the full dataset are in table 5 and the results of the subset are on page 30 in the technical appendix. In contrast, the histograms for the raters in the subset and full dataset have differing results. In particular, the histogram for the third rater using the full dataset shows that this rater favored more 2's and fewer 3’s (Figure 3). This result contrasts with the distribution of ratings for the subset of data which shows the same pattern as the other two raters (Figure 2). However, it is reassuring that the ratings for the subset of data is consistent between raters for other parts of the analysis. Between these two conclusions, the findings from the full dataset are preferred since the full dataset captures more data and the pattern of ratings are significantly different than the other two raters. Therefore, the distribution of ratings for each rater is not the same and the third rater gives lower scores.

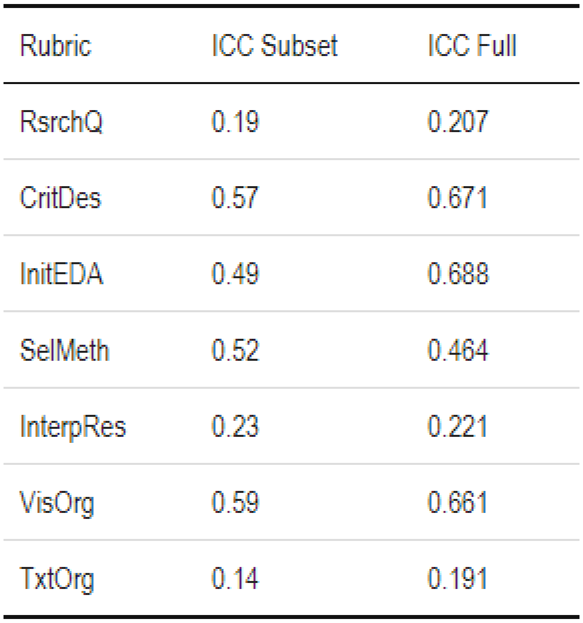
2. Rater Agreement

Recall that two methods are utilized to measure rater agreement in this study. The first method is the intra-cluster correlation (ICC) which is used to measure agreement by determining the correlation between any two rater's ratings on the same artifact. After calculating the correlations for all seven rubrics using the subset and full dataset, the ICC is moderately strong and positive for the critical design, initial exploratory data analysis, selection method, and visual organization rubrics. This result indicates that the raters had more agreement on their ratings for these rubrics. On the other hand, the raters had low correlations artifacts for the remaining rubrics, especially the text organization rubric. Therefore, there is reasonable evidence that the raters do not agree with their scores for all rubrics. These conclusions are consistent between the ratings in the subset and full datasets. The exact correlations for the full and subset of data can be found in table 6.

The other rater agreement measurement in this study is the percent of the exact agreement. From the 21 combinations of raters and rubric present in table 7, the percent of the exact agreement is comparable for the critical design, interpreting results, visual organization, and text organization rubrics. This result indicates that there was general agreement among the raters and that no rater consistently disagreed with their colleagues. However, there are large disparities in the proportions of the exact agreement for the research question, initial EDA, and selection method rubrics. A summary of the disagreements is below:

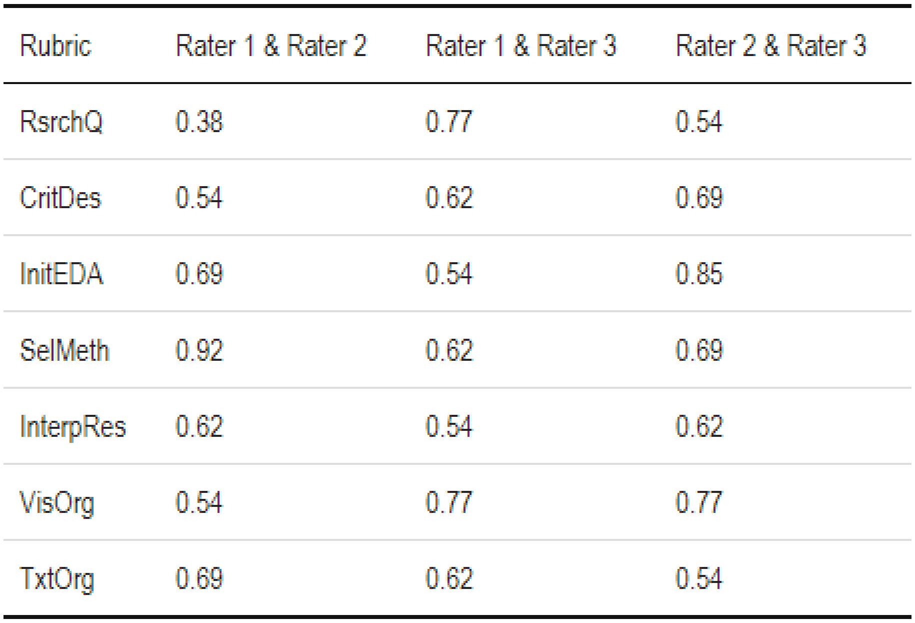
- **Research Question** – The first and second raters had little agreement and the higher proportion for the first and third rater indicates that the second rater had significantly different ratings.

Table 6: Intra-cluster correlations (ICC) for ratings in each rubric for the full dataset and the subset.



Rubric	ICC Subset	ICC Full
RsrchQ	0.19	0.207
CritDes	0.57	0.671
InitEDA	0.49	0.688
SelMeth	0.52	0.464
InterpRes	0.23	0.221
VisOrg	0.59	0.661
TxtOrg	0.14	0.191

Table 7: Percent of the exact agreement for every pairwise combination of the raters.



Rubric	Rater 1 & Rater 2	Rater 1 & Rater 3	Rater 2 & Rater 3
RsrchQ	0.38	0.77	0.54
CritDes	0.54	0.62	0.69
InitEDA	0.69	0.54	0.85
SelMeth	0.92	0.62	0.69
InterpRes	0.62	0.54	0.62
VisOrg	0.54	0.77	0.77
TxtOrg	0.69	0.62	0.54

- **Initial EDA** – The first and third rater were moderately in agreement but the higher proportion for the second and third rater indicates that the first rater had significantly different results.

The other rater agreement measurement in this study is the percent of the exact agreement. From the 21 combinations of raters and rubric present in table 7, the percent of the exact agreement is comparable for the critical design, interpreting results, visual organization, and text organization rubrics. This result indicates that there was general agreement among the raters and that no rater consistently disagreed with their colleagues. However, there are large disparities in the proportions of the exact agreement for the research question, initial EDA, and selection method rubrics. A summary of the disagreements is below:

- **Research Question** – The first and second raters had little agreement and the higher proportion for the first and third rater indicates that the second rater had significantly different ratings.
- **Initial EDA** – The first and third rater were moderately in agreement but the higher proportion for the second and third rater indicates that the first rater had significantly different results.
- **Selection Method** – The first and second raters were nearly in full agreement and the lower proportions for the other pairs of raters indicate that the third rater had significantly different ratings.

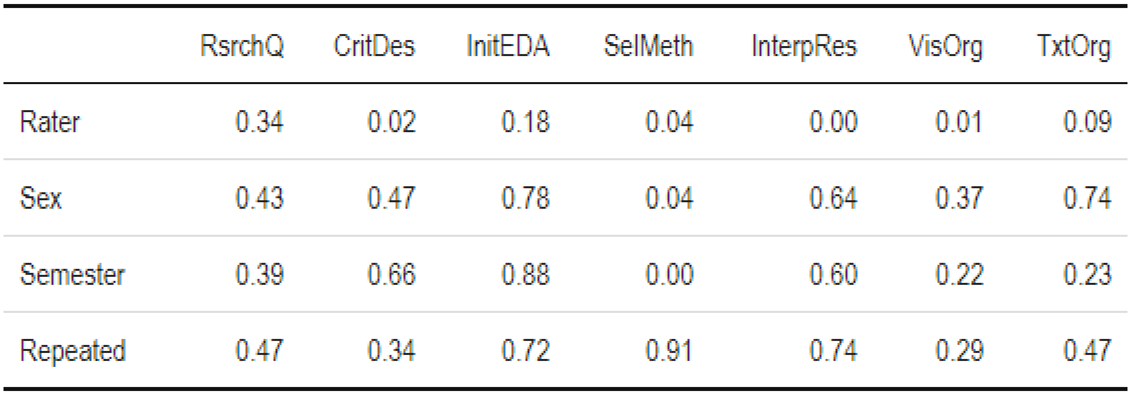
Thus, there is evidence that each rater had one rubric where their ratings were significantly different than the other two. Together, the results for the ICC and the percent of the exact agreement conclude that the raters do not generally agree on all their scores and each rater has a particular rubric where they disagreed with the other two raters.

3. Rating's Factors

Similar to the previous research question, two approaches are utilized to assess the factors that are related to the ratings. In the first technique, the significant fixed effects are added to the seven intercept-only models for each rubric. From the manual comparison, the likelihood ratio test (LRT), AIC, and BIC values from the ANOVA function had a lot of agreement (Tables 8 – 10). For one, all three metrics agree that rater should be added to the interpreting results and visual organization model, while the semester should be added to the method selection model. These are the only three variables the BIC value found to be significant to their respective models. Then, the AIC and LRT agreed on all other variable additions.

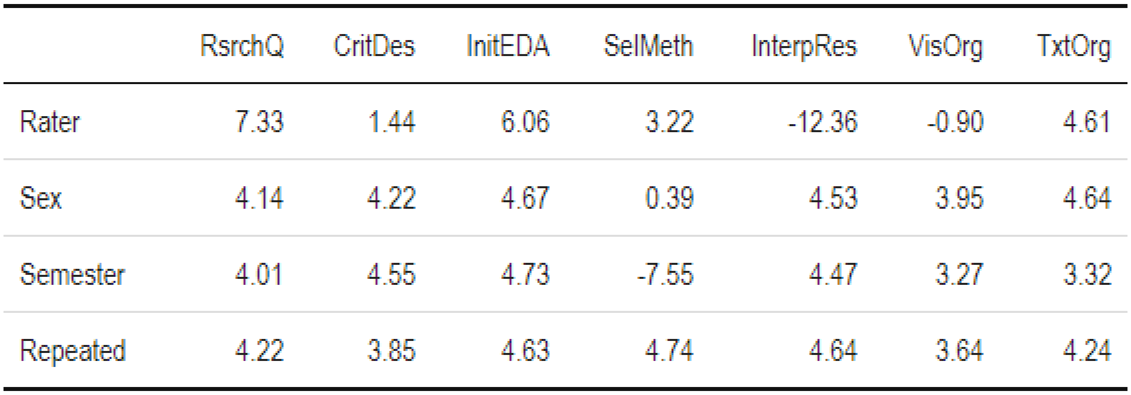
With the added effects, the selection method rubric had several fixed effects that could be added. Therefore, an ANOVA test was performed to see if all three effects needed to be included. The test resulted in a small p-value for every fixed effect except for the sex variable (Technical Appendix page 39). Therefore, the final model only contained the rater and semester as a fixed effect. The final model after this process produced the same models from the backward selection method which are presented in Table 11.

Table 8: P-values from the likelihood ratio test for every possible added fixed effect in each rubric.



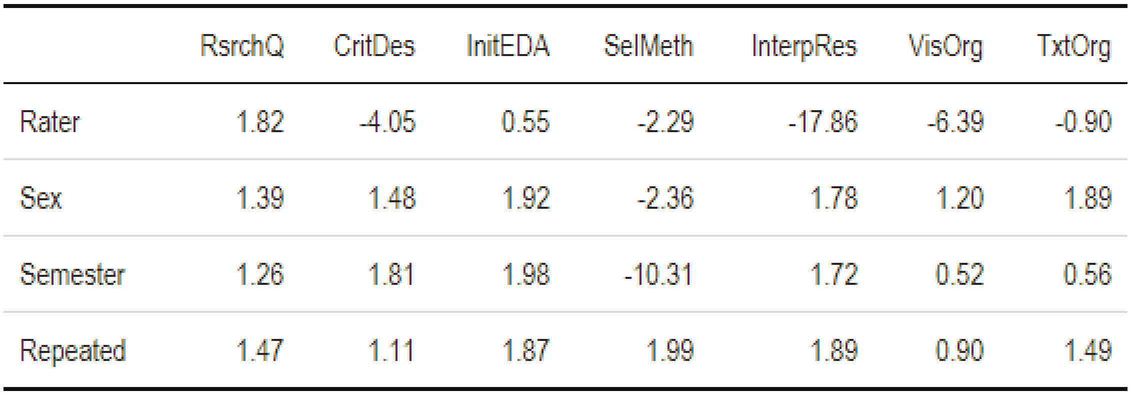
	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	0.34	0.02	0.18	0.04	0.00	0.01	0.09
Sex	0.43	0.47	0.78	0.04	0.64	0.37	0.74
Semester	0.39	0.66	0.88	0.00	0.60	0.22	0.23
Repeated	0.47	0.34	0.72	0.91	0.74	0.29	0.47

Table 9: Difference in AIC (net AIC) between the null and alternative model, positive values indicate that the model with the additional variable is worse than the null model.



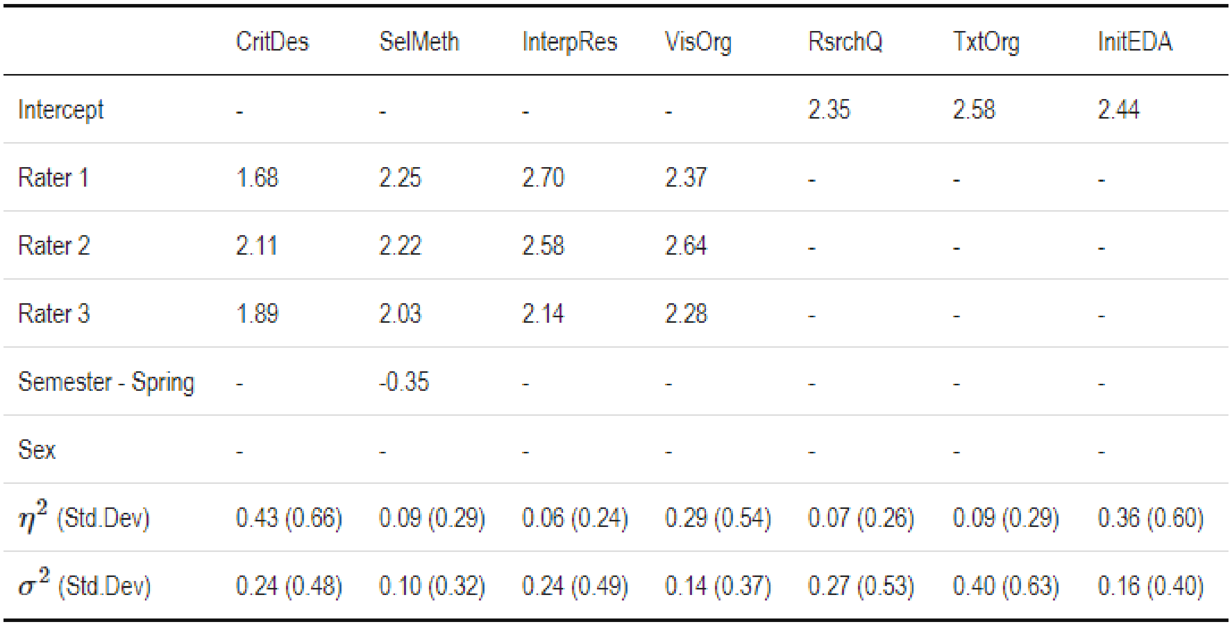
	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	7.33	1.44	6.06	3.22	-12.36	-0.90	4.61
Sex	4.14	4.22	4.67	0.39	4.53	3.95	4.64
Semester	4.01	4.55	4.73	-7.55	4.47	3.27	3.32
Repeated	4.22	3.85	4.63	4.74	4.64	3.64	4.24

Table 10: Difference in BIC (net BIC) between the null and alternative model, positive values indicate that the model with the additional variable is worse than the null model.



	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	1.82	-4.05	0.55	-2.29	-17.86	-6.39	-0.90
Sex	1.39	1.48	1.92	-2.36	1.78	1.20	1.89
Semester	1.26	1.81	1.98	-10.31	1.72	0.52	0.56
Repeated	1.47	1.11	1.87	1.99	1.89	0.90	1.49

Table 11: Summary regression statistics for the seven mixed effects models.



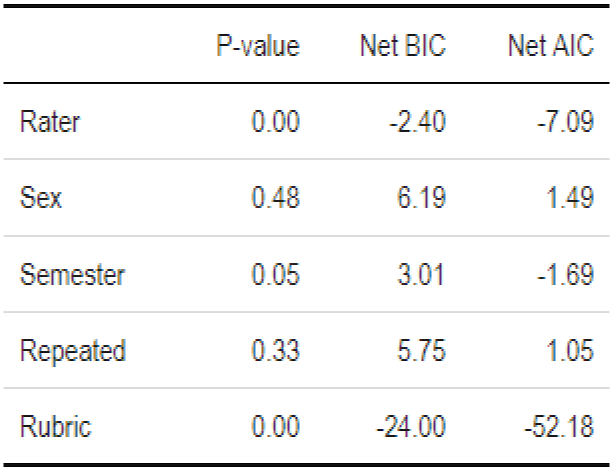
	CritDes	SelMeth	InterpRes	VisOrg	RsrchQ	TxtOrg	InitEDA
Intercept	-	-	-	-	2.35	2.58	2.44
Rater 1	1.68	2.25	2.70	2.37	-	-	-
Rater 2	2.11	2.22	2.58	2.64	-	-	-
Rater 3	1.89	2.03	2.14	2.28	-	-	-
Semester - Spring	-	-0.35	-	-	-	-	-
Sex	-	-	-	-	-	-	-
η^2 (Std.Dev)	0.43 (0.66)	0.09 (0.29)	0.06 (0.24)	0.29 (0.54)	0.07 (0.26)	0.09 (0.29)	0.36 (0.60)
σ^2 (Std.Dev)	0.24 (0.48)	0.10 (0.32)	0.24 (0.49)	0.14 (0.37)	0.27 (0.53)	0.40 (0.63)	0.16 (0.40)

After fitting the fixed effects, the selection method rubric is the only model that included more than one fixed effect. Therefore, the interaction term between the rater and semester was tested to see if it should be added to the model. The values from the ANOVA function all agreed that the interactions were not significant sense the LRT had a p-value of 0.27 and the information criterion values increased with the variable included in the model (Technical Appendix page 41). Finally, the four models that have a fixed effect were tested to see if their corresponding random effects were significant to the model. However, the rater and semester variables as random effects results in more combinations between the random effects than the number observations in the dataset. Thus, none of the random effect could be fit and the final models are still the ones presented in Table 11.

To summarize the model results, the seven mixed effect models show the various relationships that the factors have with the ratings depending on the rubric. First, the rater affects the ratings for the critical design, selection method, visual organization, and interpreting results rubric. The coefficients for the individual raters is similar within and between the seven models which does not provide meaningful insights into the significance of each rater. Additionally, the semester also effects the selection method rubric, and the negative coefficient shows that artifacts from the spring semester receive lower scores than the artifacts from the fall semesters. The remaining rubrics are not affected by the other experimental factors and are modeled by their respective average scores. This result suggests that these rubrics are a fair evaluation of the ratings since no other measured factors are affecting their scores. Lastly, it is important to note that the sex variable was insignificant to all the models which reassures that there is not an underlying gender bias with the evaluation process. Therefore, the different fixed effects between rubrics show that rater, semester, and rubric are related to the ratings.

In the other modeling technique, one model is built to test the interactions between all the fixed effects with the rubrics. All three metrics from the ANOVA function agree that rubric and rater are significant fixed effects to the model. However, AIC and the LRT suggests that the semester should also be added as a fixed effect while BIC indicates that it should be excluded. Since both AIC and p-value show that the semester should be added it is added as a fixed effect. The same final model is also the result of the backwards selection method.

Table 12: P-value, net BIC, and net AIC for all potential fixed effects on the single mixed model. For BIC and AIC, positive values indicate that the model with the additional variable is worse than the null model.

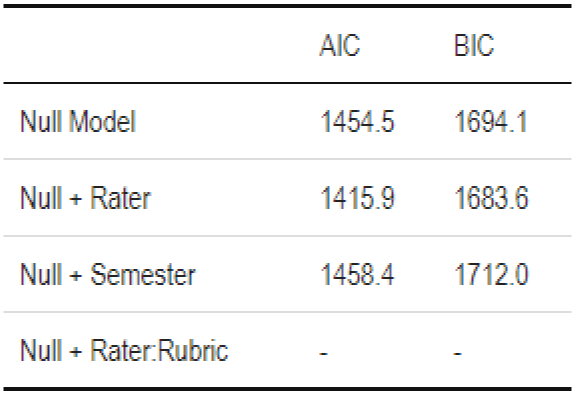


	P-value	Net BIC	Net AIC
Rater	0.00	-2.40	-7.09
Sex	0.48	6.19	1.49
Semester	0.05	3.01	-1.69
Repeated	0.33	5.75	1.05
Rubric	0.00	-24.00	-52.18

Next, all possible interaction terms from the fixed effects are evaluated to see if they should be added to the model. The backwards elimination function found that the interaction terms between rater and rubric were significant to the model. The results from the LRT in the ANOVA function agreed with this conclusion since all three metrics suggested that the model with the interactions terms was better than the model without them. Consequently, the interactions terms between rater and rubric were added to the model. Finally, the random effects for the corresponding fixed effects were assessed to see if they should be added to the model. Of the three fixed effects in the model, the rater variable as a random effect was the only variable that AIC and BIC found to be significant to the model (Table 13). Like the previous methodology, the model with the rater and rubric interactions did not produce a model since there were more random effect combinations than observations in the dataset. The final model follows the form presented in equation 1 and the coefficients for the full model can be found on page 59 in the technical appendix.

$$Rating \sim (0 + Rubric \mid Artifact) + (0 + Rater \mid Artifact) + Rater + Semester + Rubric + Rater\!:\!Rubric \qquad (1)$$

Table 13: AIC and BIC values for the models with the added random effects and the model with the final fixed effects (null model).



	AIC	BIC
Null Model	1454.5	1694.1
Null + Rater	1415.9	1683.6
Null + Semester	1458.4	1712.0
Null + Rater:Rubric	-	-

To help interpret the random effects in the final model, an overview of the variable's meaning are provided:

- **(0 + Rater | Artifact) + Rater** – There is an interaction between the raters and the artifacts. Each rater's rating on each artifact differs from the expectation by the fixed effect by a small random effect that depends on the artifact. This interaction suggests that the raters are not interpreting the artifacts the same way.
- **Rubric + Rater + Rater:Rubric** – There is an interaction between the raters and the rubric. All the raters have unique rating systems for each rubric. This interaction suggests that the raters are not interpreting the rubrics the same way.
- **(0 + Rubric | Artifact) + Rubric** – There is an interaction between the rubric and the artifacts. There are different average scores on each rubric, but the rubric averages also vary between artifacts by a small random effect that depends on the artifact. This interaction is expected since very few artifacts will contain exceptional work in all parts of the tested rubrics.

Overall, both methodologies indicate that relationships exist between the ratings, rater, semester, artifact, and rubric. The rater and semester variables are particularly important to the ratings for four of the rubrics while the semester is only significant to the selection method rubric. Then, there are interesting interactions that exist between rater, semester, rubric, and artifact outlined by the model interpretation on the previous page. These interactions show that the raters have interpretations for the rubric and artifacts.

4. Unique Rating Factors

In the previous research question, the selection method model was the only model that included the semester as a significant variable. Anticipating future questions regarding this result, the selection method model was further analyzed to investigate possible explanations. The raw counts of the artifacts per semester showed that more artifacts were sampled from the fall semester than the spring, with 87 artifacts from the fall and only 34 from the spring. Taking this disparity into account, the distribution of the ratings shows that most artifacts received a score of two in both semesters as shown in Figure 5.

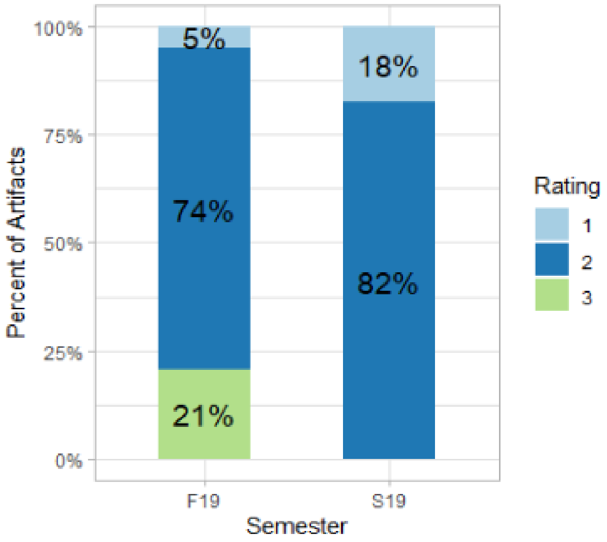


Figure 5: Proportion of ratings in the fall and spring semesters.

This observation is concerning since the inclusion of the semester variable in the final model could be linked to the homogenous distribution of ratings. Recall that the coefficient for the semester variable in the selection method model was negative, which aligns with the results in Figure 5 since more artifacts received a one and no artifact received a three. If the differences between the fall and spring semester are due to random variation, then the proportions of ratings should be similar. But the fact that the no artifact in the spring received a three when nearly a quarter of the artifacts in the fall did suggests that the disparity is not due to chance. Therefore, there is reasonable evidence that the artifacts from the spring received lower scores than the artifacts in the fall. But the cause of that disparity is not immediately obvious from the variables in the experimentation.

Discussion

In this study, the results from the experiment performed by the Dietrich college were analyzed to see what factors were contributing to the ratings given by three raters using seven different rubrics. In the first set of questions from the dean of the college, the exploratory data analysis found that the distribution of ratings varied by rubric and rater. This disparity was further proved with mixed effects models, which showed that the rater and semester were related to the rating in certain rubrics. Moreover, several interaction terms between rubric, rater, and artifact showed that the raters had unique interpretations for the rubrics and artifacts. This conclusion is particularly concerning since the rating should not be dependent upon the rater or the rubric and suggests that future iterations of the evaluation would produce different results. However, these results also point to solutions that could potentially improve the results of the ratings. For sample, the unique interpretations could be solved by additional training for the raters so that they infer the same things about the rubrics and artifacts. Further investigation of the outlying semester variable in the selection method model showed that there is a difference in ratings between the fall and spring semesters, but the cause of that disparity is not accounted for in the model. One possible explanation can be the different tracks that students take the general education courses. Or perhaps students in the fall semester are more motivated in their first semester of college compared to subsequent semesters or schedules vary based on the breaks. Overall, the results of this study have shown that the evaluation process for the artifacts in the general education courses is flawed in multiple ways. But they also indicate which areas need improvement and are immediately useful for the dean of the college.

The valuable insights are a result of the strengths within the study. For one, multiple methods were used to answer the research questions such as repeated methods for multiple parts of the dataset and using different selection techniques in the model building process. This approach allowed for a comprehensive analysis of the factors in the experimentation and additional interpretation of certain areas of interest such as rater agreement. Additionally, the use of mixed models provided valuable information about the interactions between the factors that would have been difficult to see with other tools. However, there were several limiting factors that occurred throughout the analysis. The biggest issue was with the random effects and the sample size. Future studies should sample a larger number of artifacts so that all possible random effects can be tested for significance. Moreover, the investigation of the selection method model and the interaction terms showed that there may be other factors that had an affect on the experimentation. It would be constructive for additional analyses to consider more variables that could be contributing to the ratings, such as rater background or course pedagogies.

References & Citations

Can’t think of any – any suggestions?

Technical Appendix

Data Setup

Package Import

```
library(arm)
library(lme4)
library(ggplot2)
library(plyr)
library(tidyverse)
library(reshape2)
library(kableExtra)
library(LMERConvenienceFunctions)
library(RLRsim)
```

Data Import

```
tall.data <- read.csv("C:/Users/cbrig/OneDrive/CMU/Applied Linear Models/Project 2/tall.csv")

ratings.data <- read.csv("C:/Users/cbrig/OneDrive/CMU/Applied Linear Models/Project 2/ratings.csv")
```

Data details

```
str(ratings.data)
```

```
## 'data.frame':    117 obs. of  15 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rater    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Sample   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Overlap  : int  5 7 9 8 NA NA NA NA NA 10 ...
##  $ Semester : chr  "Fall" "Fall" "Spring" "Spring" ...
##  $ Sex      : chr  "M" "F" "F" "M" ...
##  $ RsrchQ   : int  3 3 2 2 3 2 2 2 3 2 ...
##  $ CritDes  : int  3 3 1 2 3 1 1 1 1 1 ...
##  $ InitEDA  : int  2 3 3 2 3 2 3 2 2 2 ...
##  $ SelMeth  : int  2 3 2 1 3 2 2 2 2 2 ...
##  $ InterpRes: int  2 3 3 1 3 2 2 2 2 3 ...
##  $ VisOrg   : int  2 3 3 1 3 2 2 2 2 2 ...
##  $ TxtOrg   : int  3 3 3 1 3 2 2 2 2 3 ...
##  $ Artifact : chr  "05" "07" "09" "08" ...
##  $ Repeated : int  1 1 1 1 0 0 0 0 0 1 ...
```

```
dim(ratings.data)
```

```
## [1] 117  15
```

```
str(tall.data)
```

```
## ’data.frame':    819 obs. of  8 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rater   : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Artifact: chr  "O5" "O7" "O9" "O8" ...
##  $ Repeated: int  1 1 1 1 0 0 0 0 0 1 ...
##  $ Semester: chr  "F19" "F19" "S19" "S19" ...
##  $ Sex     : chr  "M" "F" "F" "M" ...
##  $ Rubric  : chr  "RsrchQ" "RsrchQ" "RsrchQ" "RsrchQ" ...
##  $ Rating  : int  3 3 2 2 3 2 2 2 3 2 ...
```

```
dim(tall.data)
```

```
## [1] 819   8
```

Initial Descriptive Statistics and EDA

```
summary(ratings.data)
```

```
##        X           Rater       Sample          Overlap     Semester
##  Min.   :  1   Min.   :1   Min.   :  1.00   Min.   : 1   Length:117
##  1st Qu.: 30   1st Qu.:1   1st Qu.: 31.00   1st Qu.: 4   Class :character
##  Median : 59   Median :2   Median : 60.00   Median : 7   Mode  :character
##  Mean   : 59   Mean   :2   Mean   : 59.89   Mean   : 7
##  3rd Qu.: 88   3rd Qu.:3   3rd Qu.: 89.00   3rd Qu.:10
##  Max.   :117   Max.   :3   Max.   :118.00   Max.   :13
##                                             NA’s   :78
##      Sex                RsrchQ        CritDes         InitEDA
##  Length:117         Min.   :1.00   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000
##  Mode  :character   Median :2.00   Median :2.000   Median :2.000
##                     Mean   :2.35   Mean   :1.871   Mean   :2.436
##                     3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000
##                     Max.   :4.00   Max.   :4.000   Max.   :4.000
##                                    NA’s   :1
##     SelMeth        InterpRes         VisOrg          TxtOrg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :2.000   Median :3.000
##  Mean   :2.068   Mean   :2.487   Mean   :2.414   Mean   :2.598
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##                                  NA’s   :1
##    Artifact            Repeated
##  Length:117         Min.   :0.0000  
##  Class :character   1st Qu.:0.0000  
##  Mode  :character   Median :0.0000  
##                     Mean   :0.3333  
##                     3rd Qu.:1.0000  
##                     Max.   :1.0000  
## 
```

Research Question 1

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

Missing Values

First, let's check to see if any of the rubrics have missing values to them.

```
# Check for NAs
colSums(is.na(ratings.data))
```

```
##         X     Rater    Sample   Overlap  Semester       Sex    RsrchQ   CritDes 
##         0         0         0        78         0         0         0         1 
##   InitEDA   SelMeth InterpRes    VisOrg    TxtOrg  Artifact  Repeated 
##         0         0         0         1         0         0         0 
```

```
# Check for other types of NA like 0 or ""
lapply(ratings.data[,c(7,8,9,10,11,12,13)], unique)
```

```
## $RsrchQ
## [1] 3 2 1 4
## 
## $CritDes
## [1]  3  1  2 NA  4
## 
## $InitEDA
## [1] 2 3 1 4
## 
## $SelMeth
## [1] 2 3 1
## 
## $InterpRes
## [1] 2 3 1 4
## 
## $VisOrg
## [1]  2  3  1  4 NA
## 
## $TxtOrg
## [1] 3 1 2 4
```

```r
# Check for NAs
colSums(is.na(tall.data))
```

```
##        X    Rater Artifact Repeated Semester      Sex   Rubric   Rating
##        0        0        0        0        0        0        0        2
```

```r
# Check for other types of NA like 0 or ""
#lapply(tall.data, unique)
```

It looks like CritDes and VisOrg have missing values, but we will check to see what these values are to see if there are any patterns.

```r
ratings.data %>% filter(is.na(CritDes))
```

```
##    X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth InterpRes
## 1 44     2     45      NA   Spring   F      2      NA       2       2         2
##   VisOrg TxtOrg Artifact Repeated
## 1      2      3       45        0
```

```r
ratings.data %>% filter(is.na(VisOrg))
```

```
##    X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth InterpRes
## 1 99     1    100      NA     Fall   F      2       3       2       3         3
##   VisOrg TxtOrg Artifact Repeated
## 1     NA      2      100        0
```

Since the missing values only occur two times and there is no obvious pattern to their missingness, we will exclude these rows from the analysis for now. There is no way that we can infer what these scores should be and it happens so infrequently that it shouldn't be considered in a separate category. Therefore, it is better to exclude the values.

```r
ratings.data <- ratings.data[!is.na(ratings.data$CritDes),]
ratings.data <- ratings.data[!is.na(ratings.data$VisOrg),]
dim(ratings.data)
```

```
## [1] 115  15
```

```r
tall.data <- tall.data[!is.na(tall.data$Rating),]
dim(tall.data)
```

```
## [1] 817   8
```

For the other predictors, it looks like there is also a missing value in the sex variable.

```r
unique(ratings.data$Sex)
```

```
## [1] "M"  "F"  "--"
```

Since we don't have enough information to infer the data, we will also exclude this value from the analysis.

```
dim(ratings.data)
```

```
## [1] 115  15
```

```
ratings.data <- ratings.data %>% filter(Sex != "--")
dim(ratings.data)
```

```
## [1] 114  15
```

```
dim(tall.data)
```

```
## [1] 817   8
```

```
tall.data <- tall.data %>% filter(Sex != "")
dim(tall.data)
```

```
## [1] 810   8
```

Now that the NAs are deleted, we can look at the distribution of the rubrics using summary statistics and box plots. Throughout the descriptive statistics, we will look at the ratings for the subset of 13 that all raters looked and the full dataset.

```
# Wide dataset
rubrics.1a.subset <- ratings.data %>% filter(Repeated == 1)
rubrics.1a.subset <- rubrics.1a.subset[,c(7,8,9,10,11,12,13)]

rubrics.1a.full <- ratings.data[,c(7,8,9,10,11,12,13)]

# Tall dataset
tall.1a.subset <- tall.data %>% filter(Repeated == 1)
tall.1a.full <- tall.data
```

Descriptive Statistics

Five number summary for subset and full dataset for every rubric:

```
# Looking at the summary statistics
#summary(rubrics.1a.subset)

apply(rubrics.1a.subset,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
# Looking at the summary statistics
#summary(rubrics.1a.full)

apply(rubrics.1a.full,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

Table 1:

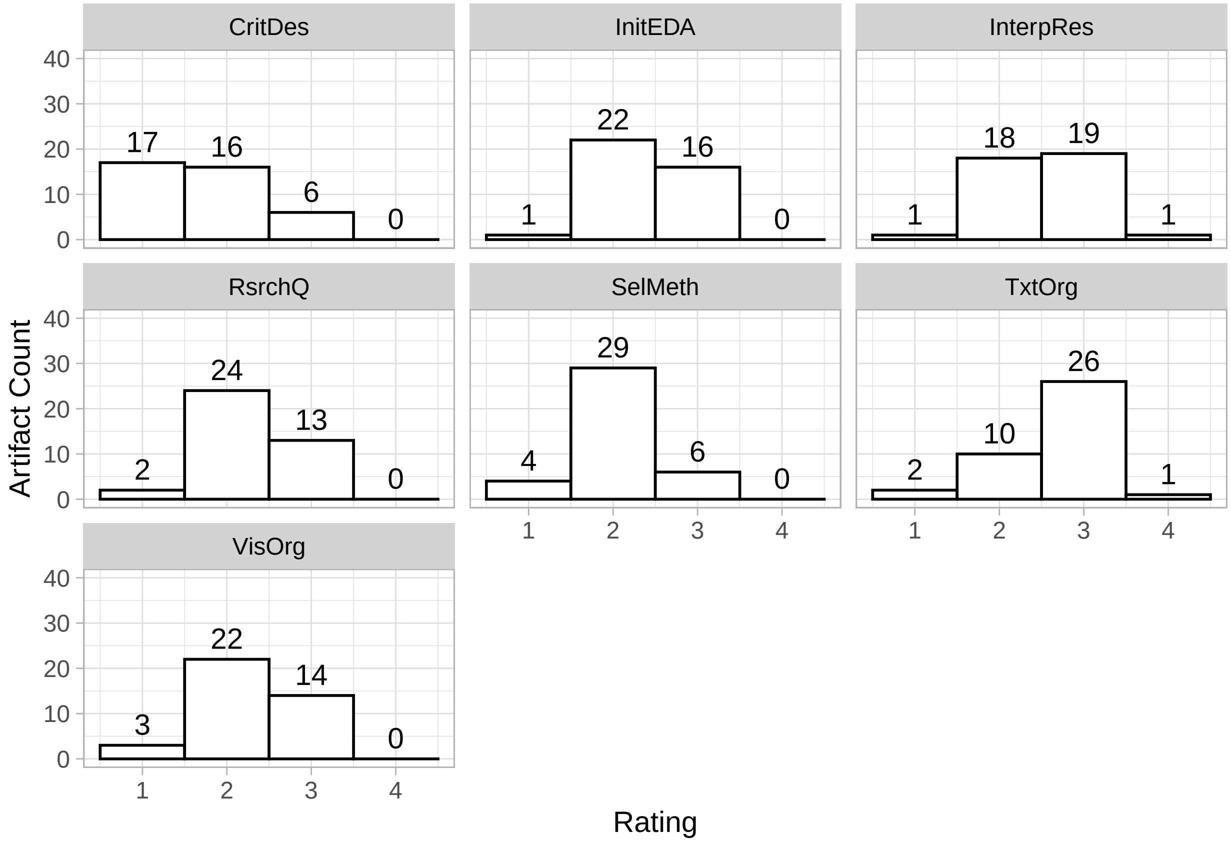
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.28	3	3	0.56
CritDes	1	1	2	1.72	2	3	0.72
InitEDA	1	2	2	2.38	3	3	0.54
SelMeth	1	2	2	2.05	2	3	0.51
InterpRes	1	2	3	2.51	3	4	0.60
VisOrg	1	2	2	2.28	3	3	0.60
TxtOrg	1	2	3	2.67	3	4	0.62

Table 2:

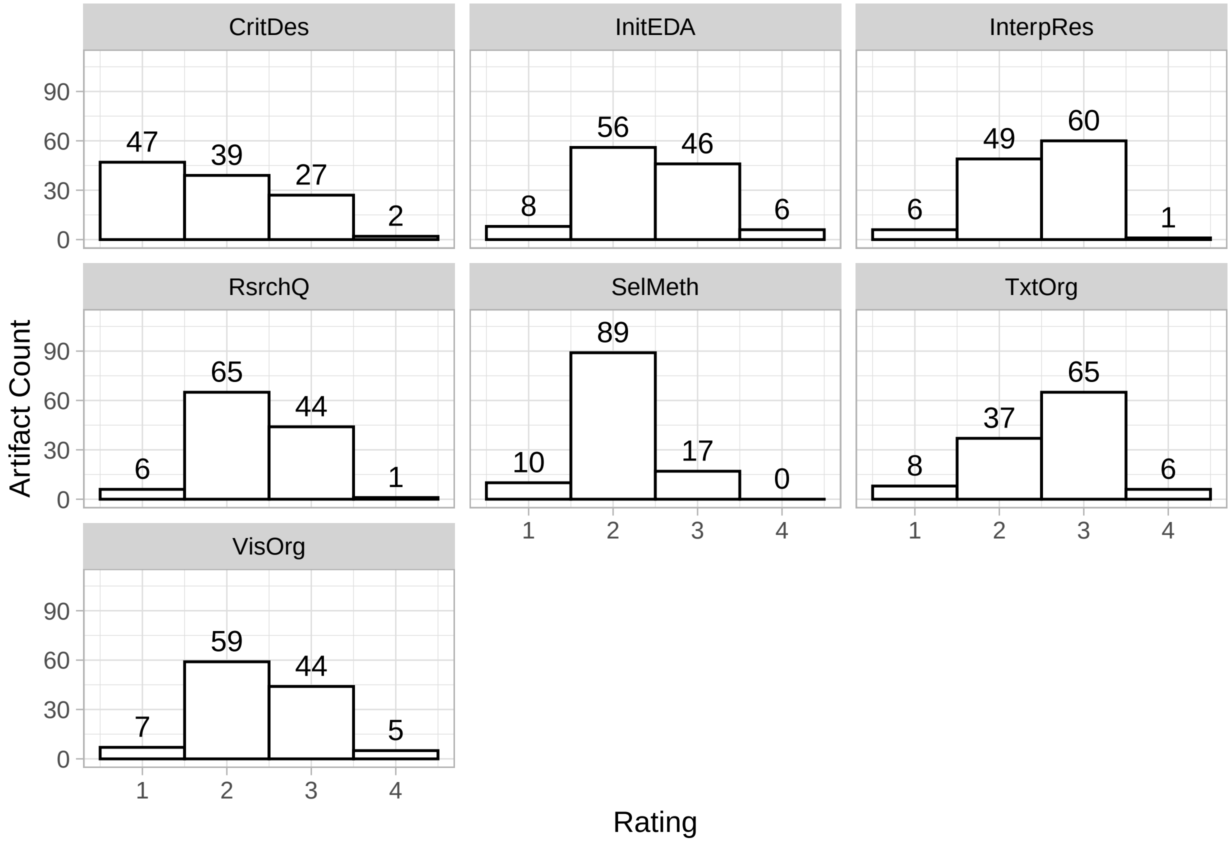
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.35	3	4	0.59
CritDes	1	1	2	1.85	2	4	0.83
InitEDA	1	2	2	2.44	3	4	0.70
SelMeth	1	2	2	2.05	2	3	0.48
InterpRes	1	2	3	2.48	3	4	0.61
VisOrg	1	2	2	2.41	3	4	0.68
TxtOrg	1	2	3	2.60	3	4	0.70

The subset and full datasets agree that there is variability in the summary statistics between the rubric. To start, the critical design rubric has more low scores since the 1st and 3rd quartile are both lower than the majority of the other rubrics and the mean is lowest among the other rubrics. Additionally, the selection method rubric seems to give a lot of 2 scores since a two was given for at least 50% of the artifacts and this rubric had the smallest standard deviation. Plus, no 4's were given for this rubric. These outlying observations suggests that the distribution of the rubrics is not homogeneous.

```
ggplot(tall.1a.subset, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1,color="black", fill="white") +
  xlab('Artifact Count')+
  scale_x_continuous(limits = c(0, 40))+
  facet_wrap(~as.factor(Rubric)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light()+
  theme(strip.background =element_rect(fill="lightgrey"))+
  theme(strip.text = element_text(colour = 'black'))
```



```r
ggplot(tall.1a.full, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1,color="black", fill="white") +
  xlab('Artifact Count')+
  scale_x_continuous(limits = c(0, 110))+
  facet_wrap(~as.factor(Rubric)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light()+
  theme(strip.background =element_rect(fill="lightgrey"))+
  theme(strip.text = element_text(colour = 'black'))
```



The histograms for the subset and full datasets agree with the conclusions from the five number summary tables. This is because the histograms also show that the critical design ratings are right skewed with the majority of artifacts receiving a 1, a trait that is unique to this rubric. Also, the histogram for the selection method rubric shows how an overwhelming majority of artifacts received a 2.

Overall, it looks like the rubrics are not the same and that the CritDes tends to give lower scores since the mean score is significantly lower than the other rubrics and the five number summary statistics are all different. These patterns are the same for the subset and full dataset so we will highlight the results from the full dataset.

Part B

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?

We will use a similar approach here with the raters as we did with the rubrics in part A: look at the summary statistics for the subset and full dataset.

```
raters.1b2 <- ratings.data %>% filter(Repeated == 1)
raters.1b2 <- raters.1b2[,c(2,7,8,9,10,11,12,13)]
# Make the table longer so that we can look at all the ratings for each rater
raters.1b2 <- gather(raters.1b2, "Rubric","Rating", 2:8)

raters.1b2 %>%
 group_by(Rater) %>%
 summarise(n = n(),
```

Table 3:

Rater	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
1	91	1	2	2	2.32	3	4	0.65
2	91	1	2	2	2.31	3	4	0.68
3	91	1	2	2	2.19	3	3	0.65

Table 4:

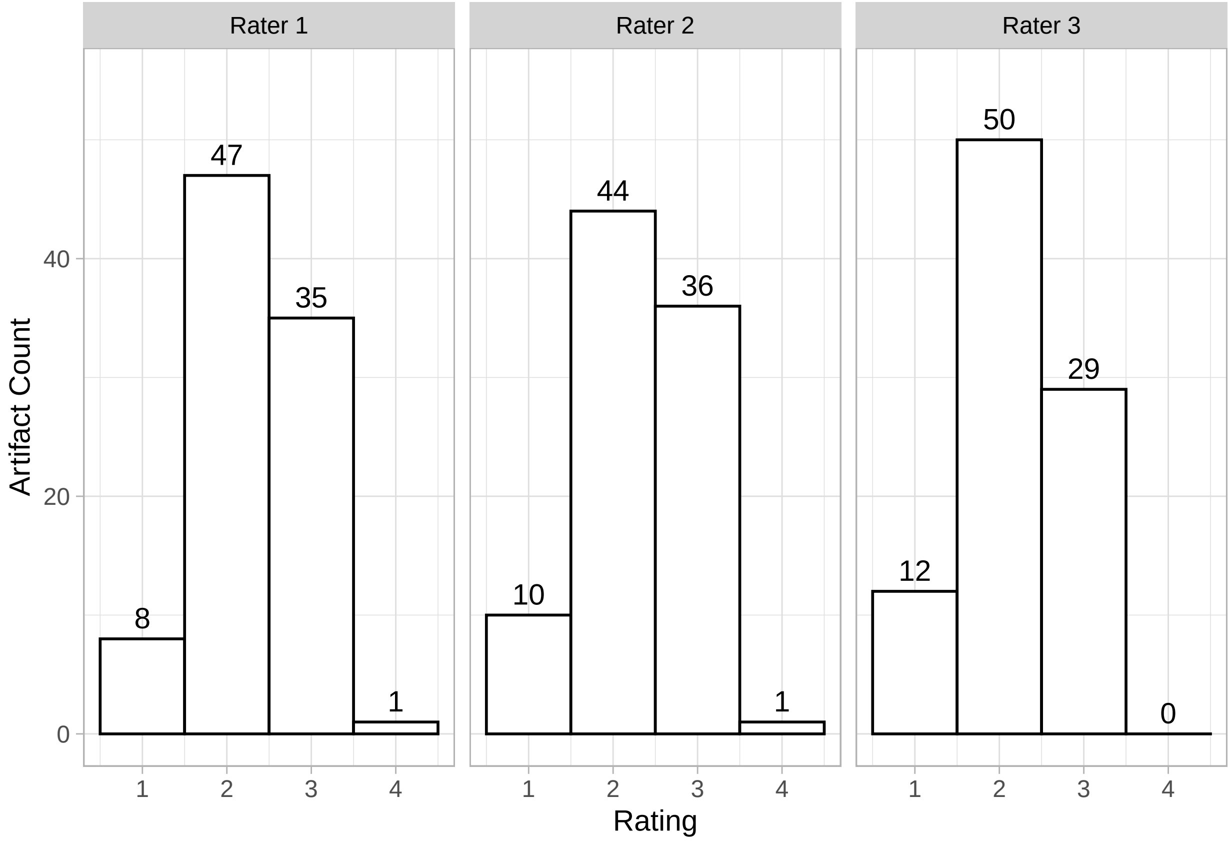
Rater	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
1	266	1	2	2	2.35	3	4	0.70
2	266	1	2	2	2.44	3	4	0.70
3	266	1	2	2	2.15	3	4	0.69

```
            Min. = fivenum(Rating)[1],
            '1st Qu.' = fivenum(Rating)[2],
            Median = fivenum(Rating)[3],
            Mean = mean(Rating),
            '3rd Qu.' = fivenum(Rating)[4],
            Max. = fivenum(Rating)[5],
            SD = sd(Rating)) %>%
  round(digits = 2) %>%
  kbl(booktabs = T, caption = " ") %>%
  kable_classic()
```

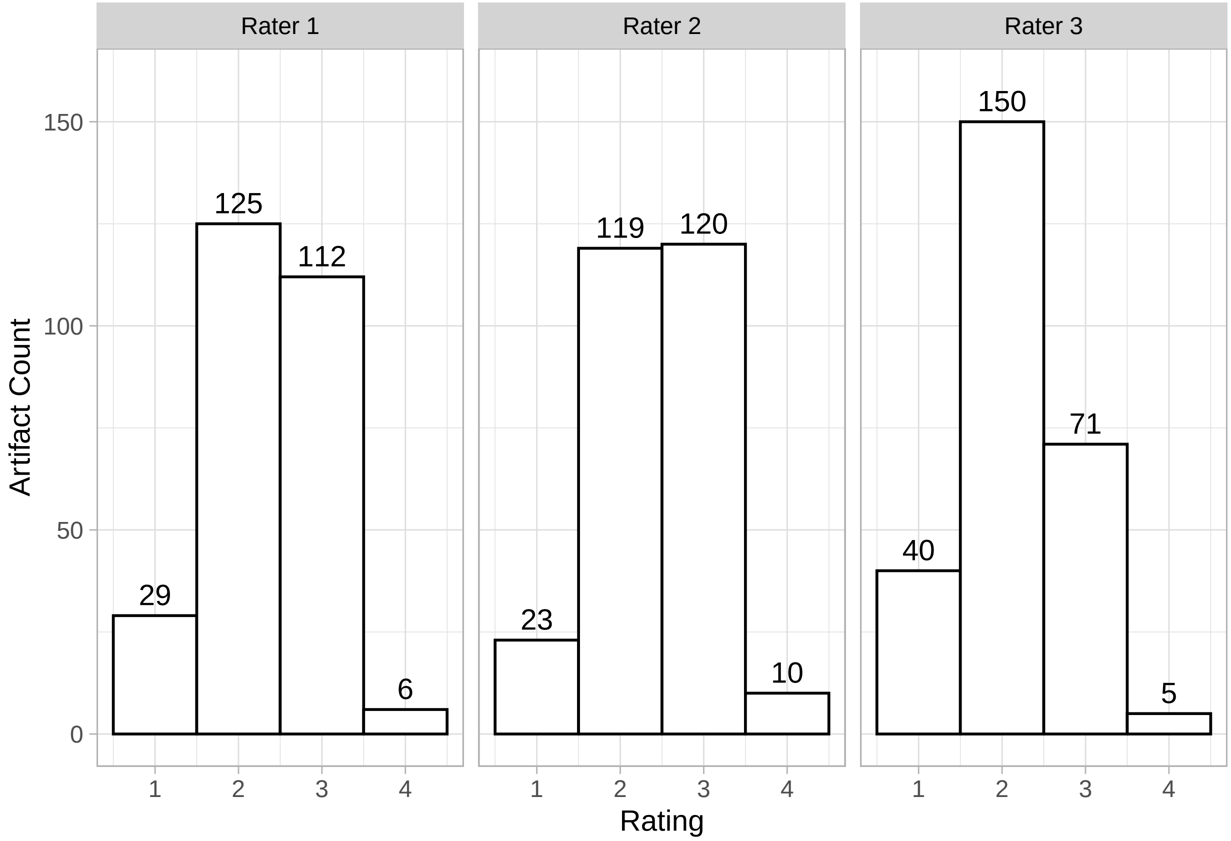
```
raters.1b <- ratings.data[,c(2,7,8,9,10,11,12,13)]

# Make the table longer so that we can look at all the ratings for each rater
raters.1b <- gather(raters.1b, "Rubric","Rating", 2:8)

raters.1b %>%
 group_by(Rater) %>%
  summarise(n = n(),
            Min. = fivenum(Rating)[1],
            '1st Qu.' = fivenum(Rating)[2],
            Median = fivenum(Rating)[3],
            Mean = mean(Rating),
            '3rd Qu.' = fivenum(Rating)[4],
            Max. = fivenum(Rating)[5],
            SD = sd(Rating)) %>%
  round(digits = 2) %>%
  kbl(booktabs = T, caption = " ") %>%
  kable_classic()
```

The summary statistics are fairly identical for the raters in the full dataset besides the lower average for rater 3. But the distribution is slightly different in the subset of data. From the subset of dataset, it looks like the distribution of ratings is nearly identical for raters 1 and 2, but rater 3 seems to give lower scores since their mean is lower and they never gave a score of 4. However, the standard deviation for all three raters is very similar. This result suggests that the distribution of ratings for each rater is not the same and that the third rater gave lower scores in comparison for the artifacts that every rater rated.

```r
tall.data.1b <- tall.data %>% filter(Repeated == 1)
rater.lable.list <- c(`1` = 'Rater 1', `2` = 'Rater 2', `3` = 'Rater 3')

ggplot(tall.data.1b, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1,color="black", fill="white") +
  xlab('Artifact Count')+
  scale_x_continuous(limits = c(0, 55))+
  facet_wrap(~as.factor(Rater), labeller = as_labeller(rater.lable.list)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light()+
  theme(strip.background =element_rect(fill="lightgrey"))+
  theme(strip.text = element_text(colour = 'black'))
```



```r
rater.lable.list <- c(`1` = 'Rater 1', `2` = 'Rater 2', `3` = 'Rater 3')

ggplot(tall.data, aes(y = Rating)) +
  geom_histogram(position = 'dodge', binwidth = 1,color="black", fill="white") +
  xlab('Artifact Count')+
  scale_x_continuous(limits = c(0, 160))+
  facet_wrap(~as.factor(Rater), labeller = as_labeller(rater.lable.list)) +
  stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-0.5) +
  coord_flip() +
  theme_light()+
  theme(strip.background =element_rect(fill="lightgrey"))+
  theme(strip.text = element_text(colour = 'black'))
```



The histograms for the raters in the subset and full dataset have similar conclusions to the summary statistics. However, the histogram for the third rater shows that this rater favored more 2's and less 3's which aligns with the conclusions from the subset of data. While the distribution of ratings for each rater looks exactly the same in the full dataset, the subset of data set shows that the third rater gives more lower scores such as two. Therefore, the distribution of rating for each rater is not the same and the third rater tended to give lower scores.

Research Question 2

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

For this question, we need to look at the artifacts where all the raters scored so that we can get an accurate comparison.

```
raters.2 <- tall.data %>%
  filter(Repeated == 1)

dim(raters.2)
```

```
## [1] 273   8
```

There are many ways to evaluate agreement, but here we will focus on three different agreement angles. We will start with looking at the intra-cluster correlation between artifacts which will allow us to investigate the association between raters. Then, we will look at exact percent agreement between the raters and ratings so that we can determine where disagreement may be occurring.

The mixed model gives seven separate models for each rubric so that we can have 3 data points for every artifact.

Part A - ICC per artifact using subset & full data

Now we will look at the ICC between any two raters' ratings on the **same** artifact. Evaluating this correlation will indicate the association between raters to see if they generally disagree or agree on the ratings.

This model will give us 13 models with three data points representing the different raters.

```r
rubric.names <- c('RsrchQ', 'CritDes', 'InitEDA', 'SelMeth', 'InterpRes', 'VisOrg', 'TxtOrg')
icc.list.subset <-  rep(NA,7)

for(i in 1:7){

  # Filter the data to the specific rubric
  rubric.ratings <- raters.2[raters.2$Rubric == rubric.names[i],]

    # Create the model
  lmer.model <- lmer(Rating ~ 1 + (1|Artifact), data = rubric.ratings)
  lmer.sum <- summary(lmer.model)

  # Find tao and sigma
  sigma <- lmer.sum$sigma^2
  tao <- as.numeric(VarCorr(lmer.model))

  # Calculating ICC
  icc <- tao/(tao + sigma)
  icc.list.subset[i] <-  icc
}
```

```r
rubric.names <- c('RsrchQ', 'CritDes', 'InitEDA', 'SelMeth', 'InterpRes', 'VisOrg', 'TxtOrg')
icc.list.full <-  rep(NA,7)

for(i in 1:7){

  # Filter the data to the specific rubric
  rubric.ratings <- tall.data[tall.data$Rubric == rubric.names[i],]

    # Create the model
  lmer.model <- lmer(Rating ~ 1 + (1|Artifact), data = rubric.ratings)
  lmer.sum <- summary(lmer.model)

  # Find tao and sigma
  sigma <- lmer.sum$sigma^2
  tao <- as.numeric(VarCorr(lmer.model))

  # Calculating ICC
  icc <- tao/(tao + sigma)
  icc.list.full[i] <-  icc
}
```

```
icc.df2b <- as.data.frame(cbind(rubric.names, round(icc.list.subset, digits = 2), round(icc.list.full, d
colnames(icc.df2b) <- c("Rubric", "ICC Subset", "ICC Full")

kableExtra::kbl(icc.df2b, caption = "", booktabs = T, linesep = "") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 5:

Rubric	ICC Subset	ICC Full
RsrchQ	0.19	0.207
CritDes	0.57	0.671
InitEDA	0.49	0.688
SelMeth	0.52	0.464
InterpRes	0.23	0.221
VisOrg	0.59	0.661
TxtOrg	0.14	0.191

Here we can see that the raters tended to agree on the CritDes, InitEDA, SelMeth, and VisOrg rubrics since they all have fairly high correlations. However, the raters disagreed on the RsrchQ, InterpRes, and TxtOrg rubrics since they all have very low correlations, especially the TxtOrg rubric.

The results for the ICC using the entire data set are very similar to the previous ICC calculations with the subset of data. Again, we can see that the raters tended to agree on the CritDes, InitEDA, SelMeth, and VisOrg rubrics since they all have fairly high correlations. However, the raters disagreed on the RsrchQ, InterpRes, and TxtOrg rubrics since they all have very low correlations, especially the TxtOrg rubric.

Part B - Exact Agreement

While the ICC can provide evidence of overall agreement, it does not tell us which raters are contributing to the disagreement. To look at disagreement, we will utilize two way tables to look at every pairwise permutation of raters with each rubric for a total of 21 tables. From these tables, we can look at the main diagonal to calculate percent exact agreement between two raters.

We will continue to look at the 13 artifacts that we were rated by all three raters but we will use the ratings table to make the calculations easier.

```
raters.2b <- ratings.data %>%
  filter(Repeated == 1)
```

```
# Keep track of the percent exact agreement
r1.r2.percent <- rep(NA,7)
r1.r3.percent <- rep(NA,7)
r2.r3.percent <- rep(NA,7)

# R1 and R2
for(i in 1:7){
  temp.df <- data.frame(
    r1 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 1],
    r2 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 2],
             a1 = raters.2b$Artifact[raters.2b$Rater == 1],
             a2 = raters.2b$Artifact[raters.2b$Rater == 2])
  r1 <- factor(temp.df$r1, levels = 1:4)
  r2 <- factor(temp.df$r2, levels = 1:4)
  tbl <- table(r1,r2)
  r1.r2.percent[i] <- sum(diag(tbl))/13
}

# R1 and R3
for(i in 1:7){
  temp.df <- data.frame(
    r1 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 1],
    r3 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 3],
             a1 = raters.2b$Artifact[raters.2b$Rater == 1],
             a2 = raters.2b$Artifact[raters.2b$Rater == 3])
  r1 <- factor(temp.df$r1, levels = 1:4)
  r3 <- factor(temp.df$r3, levels = 1:4)
  tbl <- table(r1,r3)
  r1.r3.percent[i] <- sum(diag(tbl))/13
}

# R2 and R3
for(i in 1:7){
  temp.df <- data.frame(
    r2 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 2],
    r3 = raters.2b[,colnames(raters.2b) %in% rubric.names[i]][raters.2b$Rater == 3],
             a1 = raters.2b$Artifact[raters.2b$Rater == 2],
             a2 = raters.2b$Artifact[raters.2b$Rater == 3])
  r2 <- factor(temp.df$r2, levels = 1:4)
  r3 <- factor(temp.df$r3, levels = 1:4)
  tbl <- table(r2,r3)
  r2.r3.percent[i] <- sum(diag(tbl))/13
}
```

```
percent.exact.df <- as.data.frame(cbind(rubric.names,
                    round(r1.r2.percent, digits = 2),
                    round(r1.r3.percent, digits = 2),
                    round(r2.r3.percent, digits = 2)))
colnames(percent.exact.df) <- c("Rubric", "Rater 1 & Rater 2", "Rater 1 & Rater 3", "Rater 2 & Rater 3"

kableExtra::kbl(percent.exact.df, caption = "", booktabs = T, linesep = "") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 6:

Rubric	Rater 1 & Rater 2	Rater 1 & Rater 3	Rater 2 & Rater 3
RsrchQ	0.38	0.77	0.54
CritDes	0.54	0.62	0.69
InitEDA	0.69	0.54	0.85
SelMeth	0.92	0.62	0.69
InterpRes	0.62	0.54	0.62
VisOrg	0.54	0.77	0.77
TxtOrg	0.69	0.62	0.54

Final Observations

Per Rater: > **First Rater**: Causes the most disagreement for the research question, critical design, and visual organization rubrics. > **Second rater**: Causes the most disagreement for the Initial EDA and interpreting results rubrics. > **Third rater**: Causes the most disagreement for the text organization rubric.

Per Rubric: > The research question has the most disagreement overall, with a 38% exact agreement between rater 1 and 3. > All the raters tend to agree for the remaining rubrics, with around 60% exact agreement on average. > The selection method rubric has the highest agreement overall, with rater 1 and 3 having exact agreement 92% of the time.

Combining All Parts

- **RsrchQ**: There is little correlation between the raters for this rubric and the largest disagreement occurs with the first rater.
- **CritDes**: There is positive correlation between the raters for this rubric. In other words, if one rater gives the artifact a high score the other raters are also likely to have high scores. Additionally, the raters tended to agree with their ratings.
- **InitEDA**: There is positive correlation between the raters for this rubric. In other words, if one rater gives the artifact a high score the other raters are also likely to have high scores. Additionally, the raters tended to agree with their ratings.
- **SelMeth**: There is some correlation between the raters for this rubric. In other words, if one rater gives the artifact a high score then sometimes the other raters give high scores.
- **InterpRes**: There is little correlation between the raters for this rubric but the raters tended to agree with their ratings.
- **VisOrg**: There is positive correlation between the raters for this rubric. In other words, if one rater gives the artifact a high score the other raters are also likely to have high scores. Additionally, the raters tended to agree with their ratings.
- **TxtOrg**: There is little correlation between the raters for this rubric but the raters tend to agree on what the scores are.

Research Question 3

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?

One way to do this is to add fixed effects for Rater, Semester, Sex and/or Repeated to the random intercept models for the full data set, perhaps look at interactions, and perhaps do variable selection. Do the ICC's from these models agree with your earlier ICC's? Do you find that any of these fixed effects have a significant effect in predicting ratings? Are there any other random effects that you can justify adding to these models?

We will answer these research questions by using two approaches. First, we will try to add fixed effects, interactions, and random effects to the seven models that were used in Question 2 part A. Then, we will build one model to try and answer the same questions.

Part A - Fixed effects for the seven models

We will start by trying to add fixed effects to the 7 models found in question 2 part A. For each of the models is Q2 part A, we will do variable selection to see which variables should be added to the models.

```
# Explicitly make the seven models with artifact
lmer.model.rsrch <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "RsrchQ",], R
lmer.model.crit <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "CritDes",], R
lmer.model.eda <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "InitEDA",], RE
lmer.model.sel <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "SelMeth",], RE
lmer.model.inter <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "InterpRes",]
lmer.model.visorg <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "VisOrg",],
lmer.model.txtorg <- lmer(Rating ~  1 + (1|Artifact), data = tall.data[tall.data$Rubric == "TxtOrg",],
```

```
# Visualize the modeling results with a table
# Pvalues/LRT Test
fixed.effect.pvalues.df <- as.data.frame(model.pvalues)
rownames(fixed.effect.pvalues.df) <- c('Rater','Sex',"Semester","Repeated")
colnames(fixed.effect.pvalues.df) <- rubric.names

kableExtra::kbl(fixed.effect.pvalues.df, caption = "P-values from the likelihood ratio lest for every p
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 7: P-values from the likelihood ratio lest for every possible added fixed effect in each rubric.

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	0.34	0.02	0.18	0.04	0.00	0.01	0.09
Sex	0.43	0.47	0.78	0.04	0.64	0.37	0.74
Semester	0.39	0.66	0.88	0.00	0.60	0.22	0.23
Repeated	0.47	0.34	0.72	0.91	0.74	0.29	0.47

```
# BIC
fixed.effect.bic.df <- as.data.frame(model.bic)
rownames(fixed.effect.bic.df) <- c('Rater','Sex',"Semester","Repeated")
colnames(fixed.effect.bic.df) <- rubric.names

kableExtra::kbl(fixed.effect.bic.df, caption = "Difference in BIC between null and alternative model, p
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 8: Difference in BIC between null and alternative model, positive values indicate that the model with the additional variable is worse than the null model.

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	7.33	1.44	6.06	3.22	-12.36	-0.90	4.61
Sex	4.14	4.22	4.67	0.39	4.53	3.95	4.64
Semester	4.01	4.55	4.73	-7.55	4.47	3.27	3.32
Repeated	4.22	3.85	4.63	4.74	4.64	3.64	4.24

```
# AIC
fixed.effect.aic.df <- as.data.frame(model.aic)
rownames(fixed.effect.aic.df) <- c('Rater','Sex',"Semester","Repeated")
colnames(fixed.effect.aic.df) <- rubric.names

kableExtra::kbl(fixed.effect.aic.df, caption = "Difference in AIC between null and alternative model, po
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 9: Difference in AIC between null and alternative model, positive values indicate that the model with the additional variable is worse than the null model.

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Rater	1.82	-4.05	0.55	-2.29	-17.86	-6.39	-0.90
Sex	1.39	1.48	1.92	-2.36	1.78	1.20	1.89
Semester	1.26	1.81	1.98	-10.31	1.72	0.52	0.56
Repeated	1.47	1.11	1.87	1.99	1.89	0.90	1.49

Comparing the LRT results, AIC, and BIC values, it looks like there is a lot of agreement. For one, all three metrics agree that rater should be added to the interpreting results model and semester should be added to the method selection model. These are the only two variables the BIC found to be significant to their respective models. Then, AIC and LRT agreed on all other variable additions except for two: rater to critical design and rater to initial EDA. A closer look at the difference between the AIC values show that the added variable barely decreases the AIC value. Therefore, the follow updates are made to the models:

- **RsrchQ**: No fixed
- **CritDes**: Rater
- **InitEDA**: No fixed effects
- **SelMeth**: Rater, sex, and semester are all significant fixed effects
- **InterpRes**: Rater
- **VisOrg**: Rater
- **TxtOrg**: No fixed effects

The selection method rubric is the only rubric that think several fixed effects should be added. Let's check to see if all of them are significant to the model when the other fixed effects are considered.

```
selmeth.model.full <- update(lmer.model.sel, . ~ . + as.factor(Rater) + Sex + Semester)
selmeth.model.partial <- update(lmer.model.sel, . ~ . + as.factor(Rater) + Semester)
anova(selmeth.model.partial, selmeth.model.full)
```

```
## Data: tall.data[tall.data$Rubric == "SelMeth", ]
## Models:
## selmeth.model.partial: Rating ~ (1 | Artifact) + as.factor(Rater) + Semester
## selmeth.model.full: Rating ~ (1 | Artifact) + as.factor(Rater) + Sex + Semester
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## selmeth.model.partial    6 142.05 158.58 -65.027   130.05                     
## selmeth.model.full       7 142.35 161.63 -64.178   128.35 1.6988  1     0.1924
```

Both the AIC and BIC values increased when the sex variable was also included in the model. The LRT also return a p-value that is greater than 0.05. Therefore, sex is not significant to the model when semester and rater are already included.

This leaves us with four models that included a single fixed effect

```
lmer.model.crit.updated <- update(lmer.model.crit, . ~ . + as.factor(Rater) - 1)
lmer.model.sel.updated <- update(lmer.model.sel, . ~ . + as.factor(Rater) + Semester - 1)
lmer.model.inter.updated <- update(lmer.model.inter, . ~ . + as.factor(Rater) - 1)
lmer.model.visorg.updated <- update(lmer.model.visorg, . ~ . + as.factor(Rater) - 1)
```

Part B - Significance of added fixed effects

Do you find that any of these fixed effects have a significant effect in predicting ratings?

```
writeLines('CritDes')
```

```
## CritDes
```

```
print(formula(lmer.model.crit.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.crit.updated)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 1.688428  0.1189442 14.19513
## as.factor(Rater)2 2.111667  0.1201421 17.57641
## as.factor(Rater)3 1.890745  0.1201421 15.73757
```

```
writeLines('\nSelMeth')
```

```
## 
## SelMeth
```

```
print(formula(lmer.model.sel.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) + Semester - 1
```

```
summary(lmer.model.sel.updated)$coef
```

```
##                      Estimate Std. Error  t value
## as.factor(Rater)1  2.2504313 0.07372625 30.52415
## as.factor(Rater)2  2.2265183 0.07294906 30.52155
## as.factor(Rater)3  2.0331998 0.07390189 27.51215
## SemesterS19       -0.3586506 0.09629446 -3.72452
```

```
writeLines('\nInterpRes')
```

```
## 
## InterpRes
```

```
print(formula(lmer.model.inter.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.inter.updated)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 2.703996 0.08795965 30.74132
## as.factor(Rater)2 2.585692 0.08795965 29.39635
## as.factor(Rater)3 2.139333 0.08908596 24.01425
```

```
writeLines('\nVisOrg')
```

```
## 
## VisOrg
```

```
print(formula(lmer.model.visorg.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.visorg.updated)$coef
```

```
##                   Estimate Std. Error  t value
## as.factor(Rater)1 2.376888 0.09519439 24.96878
## as.factor(Rater)2 2.648555 0.09427519 28.09387
## as.factor(Rater)3 2.285707 0.09519439 24.01094
```

Part C - Interactions

Now let's check for fixed-effect interactions. Since only one rubric has a model with at least two fixed effects, we only need to check the interactions for that model which is for the selection method rubric.

```
selmeth.interactions <- update(lmer.model.sel.updated, . ~. + as.factor(Rater)*Semester- Semester)
anova(lmer.model.sel.updated, selmeth.interactions)
```

```
## Data: tall.data[tall.data$Rubric == "SelMeth", ]
## Models:
## lmer.model.sel.updated: Rating ~ (1 | Artifact) + as.factor(Rater) + Semester - 1
## selmeth.interactions: Rating ~ (1 | Artifact) + as.factor(Rater) + as.factor(Rater):Semester - 1
##                         npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## lmer.model.sel.updated    6 142.05 158.58 -65.027   130.05
## selmeth.interactions      8 143.46 165.49 -63.731   127.46 2.592  2     0.2736
```

Looks like the fixed-effect interactions are not needed since the LRT return a pvalue that is above 0.05 and the AIC & BIC values increased when these terms were included in the model.

Part D - Variable selection using random effects

Next, we are going to rerun the variable selection process above but with random effects. Note that we will not compare the models using the p-value or the likelihood ratio test because we are comparing different random effects. We will also only focus on the models that have a fixed effect in them and see if those effects are also significant to the model as random effects.

```
# lmer.model.sel.alter <- update(lmer.model.sel.updated, . ~. + (as.factor(Rater)|Artifact))
# lmer.model.sel.new.rand <- update(lmer.model.sel.rand, . ~ . - (1|Artifact))
# exactRLRT(m0 = lmer.model.sel.updated, mA = lmer.model.sel.alter, m = lmer.model.sel.new.rand)

# Error: number of observations (=116) <= number of random effects (=270) for term
# (as.factor(Rater) | Artifact); the random-effects parameters and the residual variance (or scale
# parameter) are probably unidentifiable
```

For the alternative model, there are more random effects than there are observations in the dataset. Therefore, a mixed effects model cannot be fit and the rater cannot be a random effect

```
# lmer.model.sel.alter <- update(lmer.model.sel.updated, . ~. + (Semester|Artifact))
# lmer.model.sel.new.rand <- update(lmer.model.sel.rand, . ~ . - (1|Artifact))
# exactRLRT(m0 = lmer.model.sel.updated, mA = lmer.model.sel.alter, m = lmer.model.sel.new.rand)

# Error: number of observations (=116) <= number of random effects (=270) for term
# (as.factor(Rater) | Artifact); the random-effects parameters and the residual variance (or scale
# parameter) are probably unidentifiable
```

Again, there are more random effects than there are observations in the dataset. Therefore, a mixed effects model cannot be fit and the semester cannot be a random effect.

This means that there are no random effects in the selection method rubric. Now we will move on to the next model.

```
# lmer.model.inter.alter <- update(lmer.model.sel.updated, . ~. + (as.factor(Rater)|Artifact))
# lmer.model.inter.new.rand <- update(lmer.model.inter.rand, . ~ . - (1|Artifact))
# exactRLRT(m0 = lmer.model.inter.updated, mA = lmer.model.inter.alter, m = lmer.model.inter.new.rand)

# Error: number of observations (=116) <= number of random effects (=270) for term
# (as.factor(Rater) | Artifact); the random-effects parameters and the residual variance (or scale
# parameter) are probably unidentifiable
```

Again, there are more random effects than there are observations in the dataset. Therefore, a mixed effects model cannot be fit and the rater cannot be a random effect.

This means that there are no random effects in the interpret results rubric. Now we will move on to the next model.

```
# lmer.model.crit.alter <- update(lmer.model.crit.updated, . ~. + (as.factor(Rater)|Artifact))
# lmer.model.crit.new.rand <- update(lmer.model.crit.rand, . ~ . - (1|Artifact))
# exactRLRT(m0 = lmer.model.crit.updated, mA = lmer.model.crit.alter, m = lmer.model.crit.new.rand)

# Error: number of observations (=116) <= number of random effects (=270) for term
# (as.factor(Rater) | Artifact); the random-effects parameters and the residual variance (or scale
# parameter) are probably unidentifiable
```

Again, there are more random effects than there are observations in the dataset. Therefore, a mixed effects model cannot be fit and the rater cannot be a random effect.

This means that there are no random effects in the critical design rubric. Now we will move on to the last model.

```
# lmer.model.visorg.alter <- update(lmer.model.visorg.updated, . ~. + (as.factor(Rater)|Artifact))
# lmer.model.visorg.new.rand <- update(lmer.model.visorg.rand, . ~ . - (1|Artifact))
# exactRLRT(m0 = lmer.model.visorg.updated, mA = lmer.model.visorg.alter, m = lmer.model.visorg.new.ran

# Error: number of observations (=115) <= number of random effects (=267) for term
# (as.factor(Rater) | Artifact); the random-effects parameters and the residual variance (or scale
# parameter) are probably unidentifiable
```

Again, there are more random effects than there are observations in the dataset. Therefore, a mixed effects model cannot be fit and the rater cannot be a random effect.

This means that none of the seven models need random effects. Therefore, our final models are:

```
writeLines('CritDes')
```

```
## CritDes
```

```
print(formula(lmer.model.crit.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.crit.updated)$coef
```

```
##                    Estimate Std. Error  t value
## as.factor(Rater)1 1.688428  0.1189442 14.19513
## as.factor(Rater)2 2.111667  0.1201421 17.57641
## as.factor(Rater)3 1.890745  0.1201421 15.73757
```

```
writeLines('\nSelMeth')
```

```
##
## SelMeth
```

```
print(formula(lmer.model.sel.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) + Semester - 1
```

```
summary(lmer.model.sel.updated)$coef
```

```
##                       Estimate Std. Error  t value
## as.factor(Rater)1  2.2504313 0.07372625 30.52415
## as.factor(Rater)2  2.2265183 0.07294906 30.52155
## as.factor(Rater)3  2.0331998 0.07390189 27.51215
## SemesterS19       -0.3586506 0.09629446 -3.72452
```

```
writeLines('\nInterpRes')
```

```
##
## InterpRes
```

```
print(formula(lmer.model.inter.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.inter.updated)$coef
```

```
##                     Estimate Std. Error  t value
## as.factor(Rater)1 2.703996 0.08795965 30.74132
## as.factor(Rater)2 2.585692 0.08795965 29.39635
## as.factor(Rater)3 2.139333 0.08908596 24.01425
```

```
writeLines('\nVisOrg')
```

```
##
## VisOrg
```

```
print(formula(lmer.model.visorg.updated))
```

```
## Rating ~ (1 | Artifact) + as.factor(Rater) - 1
```

```
summary(lmer.model.visorg.updated)$coef
```

```
##                     Estimate Std. Error  t value
## as.factor(Rater)1 2.376888 0.09519439 24.96878
## as.factor(Rater)2 2.648555 0.09427519 28.09387
## as.factor(Rater)3 2.285707 0.09519439 24.01094
```

```
writeLines('\nRsrchQ')
```

```
##
## RsrchQ
```

```r
print(formula(lmer.model.rsrch))
```

```
## Rating ~ 1 + (1 | Artifact)
```

```r
summary(lmer.model.rsrch)$coef
```

```
##             Estimate Std. Error  t value
## (Intercept) 2.351432 0.05754606 40.86174
```

```r
writeLines('\nTxtOrg')
```

```
##
## TxtOrg
```

```r
print(formula(lmer.model.txtorg))
```

```
## Rating ~ 1 + (1 | Artifact)
```

```r
summary(lmer.model.txtorg)$coef
```

```
##             Estimate Std. Error  t value
## (Intercept) 2.587876 0.06768075 38.23652
```

```r
writeLines('\nInitEDA')
```

```
##
## InitEDA
```

```r
print(formula(lmer.model.eda))
```

```
## Rating ~ 1 + (1 | Artifact)
```

```r
summary(lmer.model.eda)$coef
```

```
##             Estimate Std. Error  t value
## (Intercept) 2.442222  0.0749202 32.59764
```

```r
critdes.coef <- c('-','1.68','2.11','1.89','-','-','0.43 (0.66)','0.24 (0.48)')
selmeth.coef <- c('-','2.25','2.22','2.03','-0.35','-','0.09 (0.29)','0.10 (0.32)')
interp.coef <- c('-','2.70','2.58','2.14','-','-','0.06 (0.24)','0.24 (0.49)')
visorg.coef <- c('-','2.37','2.64','2.28','-','-','0.29 (0.54)','0.14 (0.37)')
rsrchq.coef <- c('2.35','-','-','-','-','-','0.07 (0.26)','0.27 (0.53)')
txtorg.coef <- c('2.58','-','-','-','-','-','0.09 (0.29)','0.40 (0.63)')
initeda.coef <- c('2.44','-','-','-','-','-','0.36 (0.60)','0.16 (0.40)')
metrics <- c('Intercept','Rater 1', 'Rater 2', 'Rater 3', 'Semester - Spring', 'Sex', "$\\eta^2$ (Std.De
seven.model.sum <- cbind(metrics,critdes.coef,selmeth.coef,interp.coef,visorg.coef,rsrchq.coef,txtorg.co
colnames(seven.model.sum) <- c('','CritDes','SelMeth','InterpRes','VisOrg','RsrchQ','TxtOrg','InitEDA')

kableExtra::kbl(seven.model.sum, booktabs = T, linesep = "") %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

	CritDes	SelMeth	InterpRes	VisOrg	RsrchQ	TxtOrg	InitEDA
Intercept	-	-	-	-	2.35	2.58	2.44
Rater 1	1.68	2.25	2.70	2.37	-	-	-
Rater 2	2.11	2.22	2.58	2.64	-	-	-
Rater 3	1.89	2.03	2.14	2.28	-	-	-
Semester - Spring	-	-0.35	-	-	-	-	-
Sex	-	-	-	-	-	-	-
$\eta^2$ (Std.Dev)	0.43 (0.66)	0.09 (0.29)	0.06 (0.24)	0.29 (0.54)	0.07 (0.26)	0.09 (0.29)	0.36 (0.60)
$\sigma^2$ (Std.Dev)	0.24 (0.48)	0.10 (0.32)	0.24 (0.49)	0.14 (0.37)	0.27 (0.53)	0.40 (0.63)	0.16 (0.40)

Part F - Seven models interpretation

> From part A, we can conclude that the rater, semester, sex, and repeated variables do not have an association with the rating for the research question, initial eda, and visual organization rubric. But rater has a negative relationship with ratings for the selection methods, interpreting results, critical design, and text organization rubrics.
>
> Additionally, the semester and sex variables had a large effect for the ratings in the selection method rubric. But since these variables only affecting the selection method rubric, part F concludes that these two variables do not have a significant effect for all rubrics in general.
>
> From part F, we also conclude that there is an interesting interaction between rubric and rater for every artifact. Specifically part F suggests that the relationship between the rating and rater varies based on the rubric. This result aligns with our conclusions from part A since the rater is significant to a subset of the models.
>
> The repeated variable was found to be insignificant to all of the tested models.
>
> Overall, the rating depends on the rubric since the ratings have different associations between the raters, sex, and semesters within each rubric.

Part F - One model, fixed effects variable selection

This approach doesn't let you directly examine interactions with Rubric, since each model considers only one Rubric at a time (though you may find differences between the models, or in variable selection, that do suggest interactions with Rubric). One way to explore interactions with Rubric directly would be to switch to tall.csv: you might begin with the model Rating ~ (0 + Rubric | Artifact), and then add fixed effects (and possibly interactions) for all of the variables Rater, Semester, Sex, Repeated and/or Rubric, and try to answer the same kinds of questions as in the previous bullet.

We will start by fitting the intercept only model:

```
lmer.model.intercept <- lmer(Rating ~ 1 + (0 + Rubric|Artifact), data = tall.data, REML = FALSE)
summary(lmer.model.intercept)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: Rating ~ 1 + (0 + Rubric | Artifact)
##    Data: tall.data
##
##      AIC      BIC   logLik deviance df.resid
##   1527.0   1668.0   -733.5   1467.0      780
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0167 -0.4921 -0.0813  0.5249  3.7859
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.63880  0.7993
##           RubricInitEDA   0.38190  0.6180    0.25
##           RubricInterpRes 0.25549  0.5055   -0.01  0.78
##           RubricRsrchQ    0.17264  0.4155    0.38  0.50  0.74
##           RubricSelMeth   0.09484  0.3080    0.56  0.36  0.40  0.25
##           RubricTxtOrg    0.40336  0.6351    0.02  0.69  0.80  0.64  0.23
##           RubricVisOrg    0.31791  0.5638    0.17  0.78  0.77  0.59  0.28  0.79
##  Residual                 0.19449  0.4410
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.23158    0.03989   55.94
```

A boundary fit error occurs since we have strong correlations among the rubrics. Especially for the text and visual organization rubrics, which have strong correlations with each other and others. This result matches intuition since students who score well on one rubric should score well on the other rubrics. In this case, the high correlations between rubrics is not detrimental to our model so we can move forward.

Now we will try adding all the fixed effects to the model.

```
lmer.model.full <- update(lmer.model.intercept, . ~ . + as.factor(Rater) + Semester + Sex + Repeated +
summary(lmer.model.full)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Sex + Repeated + Rubric
##    Data: tall.data
##
##      AIC      BIC   logLik deviance df.resid
##   1467.5   1660.1   -692.7   1385.5      769
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1243 -0.5077 -0.0206  0.5323  3.8105
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.54169  0.7360
##           RubricInitEDA   0.34308  0.5857   0.46
##           RubricInterpRes 0.16727  0.4090   0.22 0.75
##           RubricRsrchQ    0.16227  0.4028   0.58 0.43 0.70
##           RubricSelMeth   0.06235  0.2497   0.37 0.59 0.73 0.38
##           RubricTxtOrg    0.25469  0.5047   0.33 0.61 0.70 0.55 0.66
##           RubricVisOrg    0.24923  0.4992   0.34 0.73 0.67 0.50 0.39 0.75
##  Residual                 0.18744  0.4329
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          2.014369   0.107642  18.714
## as.factor(Rater)2    0.002376   0.054326   0.044
## as.factor(Rater)3   -0.176278   0.054486  -3.235
## SemesterS19         -0.175070   0.085553  -2.046
## SexM                 0.009805   0.079127   0.124
## Repeated            -0.073472   0.095519  -0.769
## RubricInitEDA        0.547090   0.095161   5.749
## RubricInterpRes      0.587066   0.100311   5.852
## RubricRsrchQ         0.460912   0.087048   5.295
## RubricSelMeth        0.164919   0.093848   1.757
## RubricTxtOrg         0.692973   0.098936   7.004
## RubricVisOrg         0.530065   0.098583   5.377
##
## Correlation of Fixed Effects:
##             (Intr) a.(R)2 a.(R)3 SmsS19 SexM   Repetd RbIEDA RbrcIR RbrcRQ
## as.fctr(R)2 -0.245
## as.fctr(R)3 -0.238  0.499
## SemesterS19 -0.356  0.008  0.000
## SexM        -0.392 -0.026 -0.035  0.301
## Repeated    -0.152  0.001 -0.003  0.079  0.009
## RubrcIntEDA -0.556 -0.001  0.000 -0.001  0.000  0.008
## RbrcIntrpRs -0.666 -0.001  0.000 -0.001  0.000 -0.010  0.734
## RubrcRsrchQ -0.632 -0.001  0.000 -0.001  0.000 -0.040  0.585  0.756
## RubricSlMth -0.695 -0.001  0.000 -0.001  0.000 -0.089  0.658  0.776  0.689
## RubrcTxtOrg -0.616 -0.001  0.000 -0.001  0.000  0.005  0.674  0.752  0.682
## RubricVsOrg -0.612 -0.001 -0.001 -0.002 -0.001 -0.022  0.715  0.745  0.668
##             RbrcSM RbrcTO
## as.fctr(R)2
## as.fctr(R)3
## SemesterS19
## SexM
## Repeated
## RubrcIntEDA
## RbrcIntrpRs
## RubrcRsrchQ
## RubricSlMth
## RubrcTxtOrg  0.725
## RubricVsOrg  0.680  0.751
```

While no error occurs, there are still some high correlations between the rubrics. However, there are no strong correlations between the fixed effects at this point.

Manually testing the variable importance for each possible fixed effect.

```
intercept.model.pvalues <- rep(NA,5)
intercept.model.bic <- rep(NA,5)
intercept.model.aic <- rep(NA,5)

lmer.1 <- update(lmer.model.intercept, . ~ . + Rater)
lmer.2 <- update(lmer.model.intercept, . ~ . + Sex)
lmer.3 <- update(lmer.model.intercept, . ~ . + Semester)
lmer.4 <- update(lmer.model.intercept, . ~ . + Repeated)
lmer.5 <- update(lmer.model.intercept, . ~ . + Rubric)

# pvalues
intercept.model.pvalues[1] = anova(lmer.model.intercept, lmer.1)$`Pr(>Chisq)`[2]
intercept.model.pvalues[2] = anova(lmer.model.intercept, lmer.2)$`Pr(>Chisq)`[2]
intercept.model.pvalues[3] = anova(lmer.model.intercept, lmer.3)$`Pr(>Chisq)`[2]
intercept.model.pvalues[4] = anova(lmer.model.intercept, lmer.4)$`Pr(>Chisq)`[2]
intercept.model.pvalues[5] = anova(lmer.model.intercept, lmer.5)$`Pr(>Chisq)`[2]

# Compare BIC values to see if adding the random effect decreased the BIC
intercept.model.bic[1] = anova(lmer.model.intercept, lmer.1)$BIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.bic[2] = anova(lmer.model.intercept, lmer.2)$BIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.bic[3] = anova(lmer.model.intercept, lmer.3)$BIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.bic[4] = anova(lmer.model.intercept, lmer.4)$BIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.bic[5] = anova(lmer.model.intercept, lmer.5)$BIC[2] - anova(lmer.model.intercept, lmer.

# Compare BIC values to see if adding the random effect decreased the BIC
intercept.model.aic[1] = anova(lmer.model.intercept, lmer.1)$AIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.aic[2] = anova(lmer.model.intercept, lmer.2)$AIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.aic[3] = anova(lmer.model.intercept, lmer.3)$AIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.aic[4] = anova(lmer.model.intercept, lmer.4)$AIC[2] - anova(lmer.model.intercept, lmer.
intercept.model.aic[5] = anova(lmer.model.intercept, lmer.5)$AIC[2] - anova(lmer.model.intercept, lmer.
```

```
# Visualize the modeling results with a table
rand.effect.df <- as.data.frame(cbind(intercept.model.pvalues,intercept.model.bic,intercept.model.aic))
rownames(rand.effect.df) <- c('Rater','Sex',"Semester","Repeated","Rubric")
colnames(rand.effect.df) <- c("P-value","Net BIC","Net AIC")

kableExtra::kbl(rand.effect.df, caption = "", booktabs = T, linesep = "", digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 10:

	P-value	Net BIC	Net AIC
Rater	0.00	-2.40	-7.09
Sex	0.48	6.19	1.49
Semester	0.05	3.01	-1.69
Repeated	0.33	5.75	1.05
Rubric	0.00	-24.00	-52.18

The LRT, BIC, and AIC all agree that rater should be a fixed effect in the model with Rubric interacting with the artifact. This result suggests that the rating in each rubric has a different relationship for every rater. This aligns with our result from the previous parts since some models had rater has a fixed effect. In the model for this section, AIC also thinks that sex and semester should be added to the model. But since the LRT and BIC disagree with this conclusion, these two variables will be excluded from the model. Although it does suggests that sex and semester may be affecting some parts of the model which aligns with our earlier conclusions.

Now we will use the fitLMER function to backfit the fixed effects to corroborate our manual findings.

```r
lmer.model.back <- fitLMER.fnc(lmer.model.full, log.file.name = F)
```

```
## Warning in fitLMER.fnc(lmer.model.full, log.file.name = F): Argument "ran.effects" is empty, which me
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.891 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Repeated" = 0.0853 >= 0.05
##     not part of higher-order interaction

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00292562 (tol = 0.002, component 1)

##     removing term
## pruning random effects structure ...
##   nothing to prune
## ======================================================
## ===            forwardfitting random effects       ===
## ======================================================
##  ===       random slopes       ===
## ======================================================
## ===            re-backfitting fixed effects        ===
## ======================================================
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE

## boundary (singular) fit: see ?isSingular

## pruning random effects structure ...
##   nothing to prune
```

```r
summary(lmer.model.back)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Rubric
##    Data: tall.data
##
## REML criterion at convergence: 1424.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1200 -0.5125 -0.0173  0.5302  3.7752
```

```
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.55495  0.7449
##           RubricInitEDA   0.35064  0.5921   0.47
##           RubricInterpRes 0.16892  0.4110   0.23 0.75
##           RubricRsrchQ    0.16777  0.4096   0.59 0.44 0.70
##           RubricSelMeth   0.06499  0.2549   0.40 0.60 0.74 0.40
##           RubricTxtOrg    0.25615  0.5061   0.33 0.61 0.69 0.55 0.66
##           RubricVisOrg    0.25894  0.5089   0.35 0.73 0.68 0.52 0.41 0.75
##  Residual                 0.18934  0.4351
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)        2.0084130  0.0987610  20.336
## as.factor(Rater)2  0.0003231  0.0547446   0.006
## as.factor(Rater)3 -0.1771062  0.0548892  -3.227
## SemesterS19       -0.1730357  0.0826927  -2.093
## RubricInitEDA      0.5474747  0.0957148   5.720
## RubricInterpRes    0.5864544  0.1008618   5.814
## RubricRsrchQ       0.4584082  0.0874179   5.244
## RubricSelMeth      0.1590770  0.0937771   1.696
## RubricTxtOrg       0.6930033  0.0995479   6.962
## RubricVisOrg       0.5289027  0.0990973   5.337
##
## Correlation of Fixed Effects:
##             (Intr) a.(R)2 a.(R)3 SmsS19 RbIEDA RbrcIR RbrcRQ RbrcSM RbrcTO
## as.fctr(R)2 -0.281
## as.fctr(R)3 -0.277  0.499
## SemesterS19 -0.264  0.017  0.011
## RubrcIntEDA -0.610 -0.001  0.000 -0.002
## RbrcIntrpRs -0.735 -0.001  0.000  0.000  0.734
## RubrcRsrchQ -0.701 -0.001  0.000  0.002  0.586  0.756
## RubricSlMth -0.782  0.000  0.000  0.006  0.662  0.779  0.688
## RubrcTxtOrg -0.679 -0.001  0.000 -0.001  0.674  0.751  0.682  0.728
## RubricVsOrg -0.675 -0.001 -0.001  0.000  0.715  0.745  0.667  0.681  0.750
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

The selection method results are:

```
formula(lmer.model.back)
```

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

The final model includes the rater, semester, and rubric as fixed effects which aligns with our conclusions from the manual calculations.

Even though we get a boundary error, we are still going to check if any interactions need to be added to the model.

```
lmer.model.interact <- update(lmer.model.back, .~. + as.factor(Rater)*Semester*Rubric)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00371227 (tol = 0.002, component 1)
```

Including the interactions causes a convergence warning. To avoid this, we will use a different optimizer.

```
ss <- getME(lmer.model.interact, c("theta", "fixef"))
lmer.model.interact.updated <- update(lmer.model.interact, start = ss, control = lmerControl(optimizer
```

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

```
summary(lmer.model.interact.updated)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Rubric + as.factor(Rater):Semester + as.factor(Rater):Rubric +
##     Semester:Rubric + as.factor(Rater):Semester:Rubric
##    Data: tall.data
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1424.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9141 -0.5141 -0.0653  0.5023  3.6609
##
## Random effects:
##  Groups   Name             Variance Std.Dev. Corr
##  Artifact RubricCritDes    0.48550  0.6968
##           RubricInitEDA    0.35257  0.5938   0.42
##           RubricInterpRes  0.14619  0.3824   0.32 0.80
##           RubricRsrchQ     0.16444  0.4055   0.66 0.43 0.72
##           RubricSelMeth    0.06297  0.2509   0.45 0.64 0.78 0.49
##           RubricTxtOrg     0.25441  0.5044   0.44 0.65 0.67 0.60 0.62
##           RubricVisOrg     0.25527  0.5052   0.35 0.73 0.68 0.57 0.35 0.76
##  Residual                  0.18839  0.4340
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                                 Estimate Std. Error t value
## (Intercept)                                     1.739538   0.136568  12.738
## as.factor(Rater)2                               0.302995   0.155107   1.953
## as.factor(Rater)3                               0.237851   0.155863   1.526
## SemesterS19                                    -0.129077   0.250318  -0.516
## RubricInitEDA                                   0.765215   0.165241   4.631
## RubricInterpRes                                 0.979228   0.162160   6.039
## RubricRsrchQ                                    0.710427   0.147386   4.820
## RubricSelMeth                                   0.462750   0.155274   2.980
## RubricTxtOrg                                    1.011251   0.160899   6.285
## RubricVisOrg                                    0.647869   0.166603   3.889
## as.factor(Rater)2:SemesterS19                  0.268014   0.303883   0.882
## as.factor(Rater)3:SemesterS19                 -0.072789   0.301026  -0.242
## as.factor(Rater)2:RubricInitEDA               -0.325018   0.204108  -1.592
## as.factor(Rater)3:RubricInitEDA               -0.374190   0.205354  -1.822
## as.factor(Rater)2:RubricInterpRes             -0.469281   0.201051  -2.334
## as.factor(Rater)3:RubricInterpRes             -0.711515   0.202316  -3.517
## as.factor(Rater)2:RubricRsrchQ                -0.447050   0.189326  -2.361
## as.factor(Rater)3:RubricRsrchQ                -0.474411   0.190681  -2.488
## as.factor(Rater)2:RubricSelMeth               -0.301450   0.193678  -1.556
## as.factor(Rater)3:RubricSelMeth               -0.365656   0.194970  -1.875
## as.factor(Rater)2:RubricTxtOrg                -0.449164   0.200927  -2.235
## as.factor(Rater)3:RubricTxtOrg                -0.407754   0.202209  -2.016
## as.factor(Rater)2:RubricVisOrg                 0.009042   0.205059   0.044
## as.factor(Rater)3:RubricVisOrg                -0.287443   0.206299  -1.393
## SemesterS19:RubricInitEDA                     -0.050212   0.301475  -0.167
## SemesterS19:RubricInterpRes                    0.127813   0.295706   0.432
## SemesterS19:RubricRsrchQ                       0.133874   0.267750   0.500
## SemesterS19:RubricSelMeth                     -0.089616   0.282837  -0.317
## SemesterS19:RubricTxtOrg                       0.166097   0.293176   0.567
## SemesterS19:RubricVisOrg                       0.146845   0.302496   0.485
## as.factor(Rater)2:SemesterS19:RubricInitEDA    0.020326   0.392376   0.052
## as.factor(Rater)3:SemesterS19:RubricInitEDA    0.252422   0.389961   0.647
## as.factor(Rater)2:SemesterS19:RubricInterpRes -0.266618   0.385390  -0.692
## as.factor(Rater)3:SemesterS19:RubricInterpRes -0.152392   0.383354  -0.398
## as.factor(Rater)2:SemesterS19:RubricRsrchQ    -0.217348   0.360414  -0.603
## as.factor(Rater)3:SemesterS19:RubricRsrchQ     0.354319   0.357388   0.991
## as.factor(Rater)2:SemesterS19:RubricSelMeth   -0.401036   0.370200  -1.083
## as.factor(Rater)3:SemesterS19:RubricSelMeth   -0.192670   0.367887  -0.524
## as.factor(Rater)2:SemesterS19:RubricTxtOrg    -0.542267   0.385011  -1.408
## as.factor(Rater)3:SemesterS19:RubricTxtOrg    -0.316395   0.382614  -0.827
## as.factor(Rater)2:SemesterS19:RubricVisOrg    -0.603626   0.392909  -1.536
## as.factor(Rater)3:SemesterS19:RubricVisOrg    -0.186749   0.390759  -0.478
```

```
##
## Correlation matrix not shown by default, as p = 42 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

Now that we have fitted the interaction terms, we will use the backward selection function again to determine which interaction terms should be added to the model.

```
lmer.model.interact.back <- fitLMER.fnc(lmer.model.interact.updated, log.file.name = F)
```

```
## Warning in fitLMER.fnc(lmer.model.interact.updated, log.file.name = F): Argument "ran.effects" is emp
## TRUE
```

```
## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 3
##   iteration 1
##     p-value for term "as.factor(Rater):Semester:Rubric" = 0.5526 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
## processing model terms of interaction level 2
##   iteration 2
##     p-value for term "as.factor(Rater):Semester" = 0.598 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
##   iteration 3
##     p-value for term "Semester:Rubric" = 0.0761 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     removing term
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## pruning random effects structure ...
##   nothing to prune
## ======================================================
## ===            forwardfitting random effects       ===
## ======================================================
##  ===       random slopes       ===
## ======================================================
## ===            re-backfitting fixed effects        ===
## ======================================================
## processing model terms of interaction level 2
##   all terms of interaction level 2 significant
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE

## boundary (singular) fit: see ?isSingular

## pruning random effects structure ...
##   nothing to prune
```

```
summary(lmer.model.interact.back)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester +
##     Rubric + as.factor(Rater):Rubric
##    Data: tall.data
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
## REML criterion at convergence: 1419.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9280 -0.5122 -0.0447  0.4827  3.5854
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.50348  0.7096
##           RubricInitEDA   0.35480  0.5956   0.44
##           RubricInterpRes 0.15192  0.3898   0.35 0.82
##           RubricRsrchQ    0.17953  0.4237   0.63 0.44 0.72
##           RubricSelMeth   0.06727  0.2594   0.42 0.60 0.74 0.36
##           RubricTxtOrg    0.26069  0.5106   0.42 0.64 0.67 0.55 0.64
##           RubricVisOrg    0.25491  0.5049   0.34 0.71 0.68 0.51 0.38 0.77
##  Residual                 0.18519  0.4303
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                    Estimate Std. Error t value
## (Intercept)                         1.75945    0.11785  14.929
## as.factor(Rater)2                   0.36537    0.13296   2.748
## as.factor(Rater)3                   0.21421    0.13297   1.611
## SemesterS19                        -0.17780    0.08228  -2.161
## RubricInitEDA                       0.74625    0.13676   5.457
## RubricInterpRes                     1.01453    0.13479   7.527
## RubricRsrchQ                        0.74926    0.12419   6.033
## RubricSelMeth                       0.42672    0.13040   3.272
## RubricTxtOrg                        1.04967    0.13551   7.746
## RubricVisOrg                        0.68354    0.13947   4.901
## as.factor(Rater)2:RubricInitEDA   -0.30843    0.17249  -1.788
## as.factor(Rater)3:RubricInitEDA   -0.29522    0.17282  -1.708
## as.factor(Rater)2:RubricInterpRes -0.53674    0.17008  -3.156
## as.factor(Rater)3:RubricInterpRes -0.75247    0.17049  -4.414
## as.factor(Rater)2:RubricRsrchQ    -0.50157    0.16151  -3.106
## as.factor(Rater)3:RubricRsrchQ    -0.37068    0.16179  -2.291
## as.factor(Rater)2:RubricSelMeth   -0.39602    0.16467  -2.405
## as.factor(Rater)3:RubricSelMeth   -0.41324    0.16504  -2.504
## as.factor(Rater)2:RubricTxtOrg    -0.58380    0.17141  -3.406
## as.factor(Rater)3:RubricTxtOrg    -0.48649    0.17177  -2.832
## as.factor(Rater)2:RubricVisOrg    -0.14444    0.17442  -0.828
## as.factor(Rater)3:RubricVisOrg    -0.33380    0.17481  -1.910
```

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

```
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

The selection method results are:

```
formula(lmer.model.interact.back)
```

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

This model contains the rater, semester, and rubric as the fixed effects with the rater and rubric interaction terms.

Now we use ANOVA to check if the interaction terms should be included holistically by comparing the model with some of the previous findings.

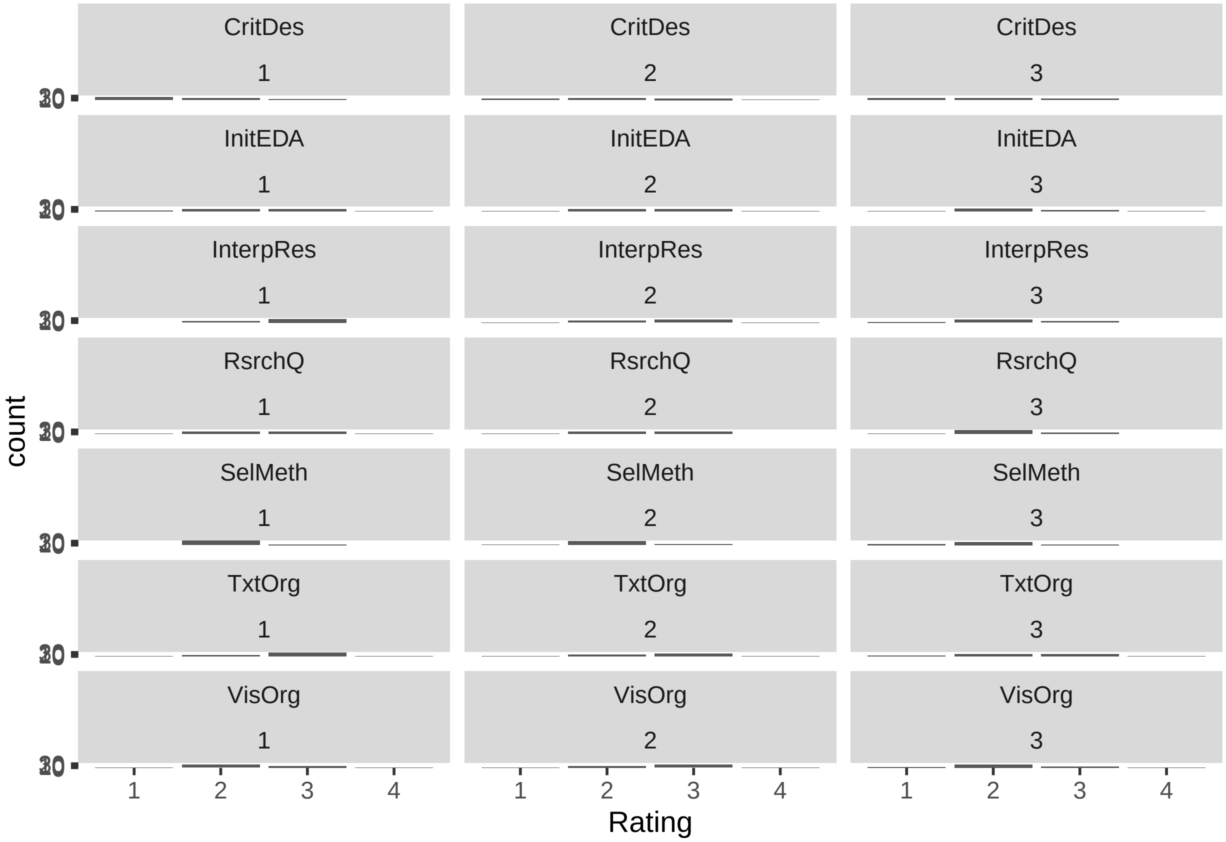
```
anova(lmer.model.back, lmer.model.interact.back, lmer.model.interact.updated)
```

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

```
## Data: tall.data
## Models:
## lmer.model.back: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric
## lmer.model.interact.back: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + a
## lmer.model.interact.updated: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric
##                             npar    AIC    BIC  logLik deviance  Chisq Df
## lmer.model.back               39 1464.0 1647.2 -693.02   1386.0
## lmer.model.interact.back      51 1454.5 1694.1 -676.26   1352.5 33.526 12
## lmer.model.interact.updated   71 1471.4 1804.8 -664.68   1329.4 23.161 20
##                             Pr(>Chisq)
## lmer.model.back
## lmer.model.interact.back      0.000801 ***
## lmer.model.interact.updated   0.280962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test shows that AIC prefers the backward selected interactions model while BIC prefers the backward selection model without any interaction terms. All metrics agree that the model with all the interaction terms is not the best model. The model suggested by AIC has interactions between rubric and rater which says that raters tended to give different ratings between each rubric. Let's see if this is true using histograms:

```
ggplot(tall.data, aes(x=Rating)) +
  geom_bar() +
  facet_wrap( ~ Rubric + Rater, nrow=7)
```



It does look like raters gave different ratings depending on the rubric. Therefore, we will use the final model from the backward elimination process with the interaction terms.

Finally, we will try to add random effects to the model. We have three sets of fixed effects so we will see if their corresponding random effects are significant to the model using ANOVA.

First we will see if Rater should be added to the model as a random effect:

```
lmer.model.rand1 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifac
```

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

```
anova(lmer.model.interact.back, lmer.model.rand1)
```

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

```
## Data: tall.data
## Models:
## lmer.model.interact.back: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + a
## lmer.model.rand1: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) +
##                          npar    AIC    BIC  logLik deviance  Chisq Df
## lmer.model.interact.back   51 1454.5 1694.1 -676.26   1352.5
## lmer.model.rand1           57 1415.9 1683.6 -650.94   1301.9 50.647  6
##                          Pr(>Chisq)
## lmer.model.interact.back
## lmer.model.rand1          3.487e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AIC, BIC, and LRT all agree that the random effect for rater should be included in the model so we will include it in the final model.

Next we will test to see if Semester should be added as a random effect:

```
lmer.model.rand2 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + as.
```

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

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

```
anova(lmer.model.interact.back, lmer.model.rand2)
```

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

```
## Data: tall.data
## Models:
## lmer.model.interact.back: Rating ~ (0 + Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + a
## lmer.model.rand2: as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + Semester | Artifact) + as.facto
##                          npar    AIC    BIC  logLik deviance  Chisq Df
## lmer.model.interact.back   51 1454.5 1694.1 -676.26   1352.5
## lmer.model.rand2           54 1458.4 1712.0 -675.18   1350.4 2.1534  3
##                          Pr(>Chisq)
## lmer.model.interact.back
## lmer.model.rand2             0.5412
```

AIC, BIC, and LRT all agree that the random effect for semester should not be included in the model so we won't include it in the final model.

Lastly, we will see if the interaction terms should be included as a random effect:

```
# lmer.model.rand3 <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artif
# (0 + as.factor(Rater):Rubric | Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Ru
#
# Error: number of observations (=810) <= number of random effects (=1890) for term (0 + as.factor(Rate
```

We get the same error as we did in the previous models which means that the random effect for the interaction terms should not be included in the final model so we won't include them.

```
rand.aic <- c("1454.5", "1415.9", "1458.4", "-")
rand.bic <- c("1694.1", "1683.6", '1712.0', "-")
rand.effects <- c("Null Model", "Null + Rater", "Null + Semester", "Null + Rater:Rubric")
rand.effects.sum <- cbind(rand.effects, rand.aic, rand.bic)
colnames(rand.effects.sum) <- c('','AIC','BIC')

kableExtra::kbl(rand.effects.sum, caption = "", booktabs = T, linesep = "", digits = 2) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 11:

	AIC	BIC
Null Model	1454.5	1694.1
Null + Rater	1415.9	1683.6
Null + Semester	1458.4	1712.0
Null + Rater:Rubric	-	-

This leaves us with the final model:

```
lmer.final <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) | Artifact) + as
Semester + Rubric + as.factor(Rater):Rubric, data = tall.data)
```

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

```
formula(lmer.final)
```

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

```
summary(lmer.final)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## as.numeric(Rating) ~ (0 + Rubric | Artifact) + (0 + as.factor(Rater) |
##     Artifact) + as.factor(Rater) + Semester + Rubric + as.factor(Rater):Rubric
##    Data: tall.data
##
## REML criterion at convergence: 1370.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.06428 -0.46900 -0.02983  0.45341  2.74000
##
## Random effects:
##  Groups     Name              Variance Std.Dev. Corr
##  Artifact   RubricCritDes     0.49642  0.7046
##             RubricInitEDA     0.31786  0.5638    0.32
##             RubricInterpRes   0.10206  0.3195    0.14  0.67
##             RubricRsrchQ      0.17899  0.4231    0.50  0.19  0.54
##             RubricSelMeth     0.03824  0.1956    0.14  0.23  0.38 -0.24
##             RubricTxtOrg      0.25028  0.5003    0.27  0.44  0.36  0.31  0.21
##             RubricVisOrg      0.23234  0.4820    0.18  0.50  0.45  0.28 -0.16
##  Artifact.1 as.factor(Rater)1 0.01281  0.1132
##             as.factor(Rater)2 0.11175  0.3343   -0.49
##             as.factor(Rater)3 0.09414  0.3068    0.33  0.66
##  Residual                     0.13468  0.3670
##
##
##
```

```
##
##
##
##
##   0.54
##
##
##
##
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                                   Estimate Std. Error t value
## (Intercept)                        1.75755    0.11404  15.412
## as.factor(Rater)2                  0.36606    0.13918   2.630
## as.factor(Rater)3                  0.19591    0.12967   1.511
## SemesterS19                       -0.15917    0.07647  -2.081
## RubricInitEDA                      0.73950    0.12996   5.690
## RubricInterpRes                    0.99152    0.12771   7.764
## RubricRsrchQ                       0.72619    0.11793   6.158
## RubricSelMeth                      0.41068    0.12470   3.293
## RubricTxtOrg                       1.01578    0.13000   7.814
## RubricVisOrg                       0.65425    0.13353   4.900
## as.factor(Rater)2:RubricInitEDA   -0.29981    0.15609  -1.921
## as.factor(Rater)3:RubricInitEDA   -0.29473    0.15635  -1.885
## as.factor(Rater)2:RubricInterpRes -0.51324    0.15348  -3.344
## as.factor(Rater)3:RubricInterpRes -0.71484    0.15364  -4.653
## as.factor(Rater)2:RubricRsrchQ    -0.48741    0.14722  -3.311
## as.factor(Rater)3:RubricRsrchQ    -0.32238    0.14727  -2.189
## as.factor(Rater)2:RubricSelMeth   -0.38638    0.15031  -2.571
## as.factor(Rater)3:RubricSelMeth   -0.38716    0.14961  -2.588
## as.factor(Rater)2:RubricTxtOrg    -0.55105    0.15646  -3.522
## as.factor(Rater)3:RubricTxtOrg    -0.44488    0.15673  -2.839
## as.factor(Rater)2:RubricVisOrg    -0.10490    0.15861  -0.661
## as.factor(Rater)3:RubricVisOrg    -0.27521    0.15885  -1.733
```

```
##
## 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
```

Research Question 4

Is there anything else interesting to say about this data?

One of the results that stood out was the inclusion of sex and semester in the mixed effects model for the selection method rubric. Here we will look further into this result to see why these variables were added to the model.

First we will only look at the data for the selection method rubric.

```
sel.method.data <- tall.data %>% filter(Rubric == 'SelMeth')
head(sel.method.data)
```

```
##     X Rater Artifact Repeated Semester Sex  Rubric Rating
## 1 352     3       05        1      F19   M SelMeth      2
## 2 353     3       07        1      F19   F SelMeth      3
## 3 354     3       09        1      S19   F SelMeth      2
## 4 355     3       08        1      S19   M SelMeth      1
## 5 357     3        6        0      F19   M SelMeth      2
## 6 358     3        7        0      F19   F SelMeth      2
```

We will start by looking at rating by semester and rating.

```
q4.viz.data <- sel.method.data %>%
  group_by(Semester, Rating) %>%
  summarize( n = n())
```

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

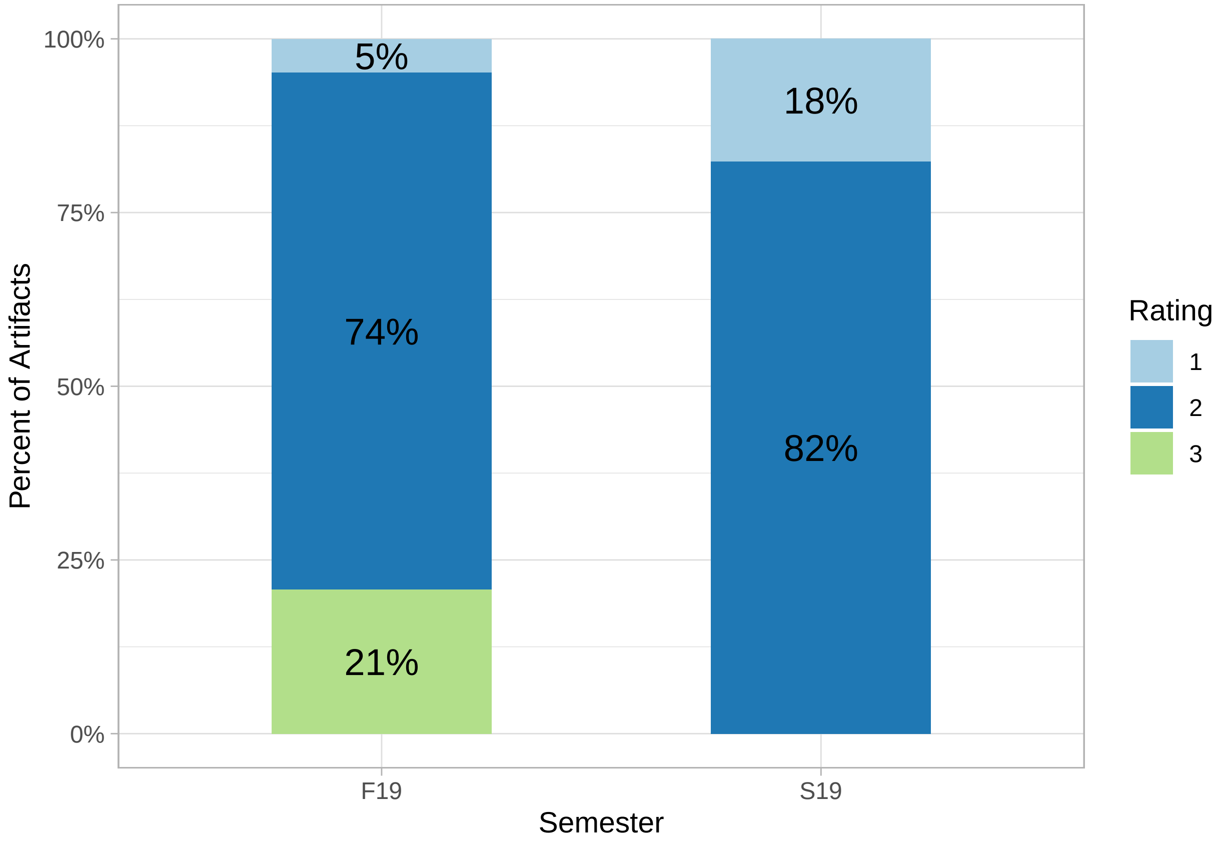
```
q4.viz.data <- cbind(q4.viz.data, c(82,82,82,34,34))
```

```
## New names:
## * NA -> ...4
```

```
q4.viz.data <- q4.viz.data %>% mutate(n = n / `...4`)

head(q4.viz.data)
```

```
## # A tibble: 5 x 4
## # Groups:   Semester [2]
##   Semester Rating      n  ...4
##   <chr>    <int>  <dbl> <dbl>
## 1 F19           1 0.0488    82
## 2 F19           2 0.744     82
## 3 F19           3 0.207     82
## 4 S19           1 0.176     34
## 5 S19           2 0.824     34
```

```
ggplot(q4.viz.data, aes(x = Semester, y = n, fill = factor(Rating))) +
  geom_bar(position="fill",stat="identity",width=0.5, ) +
  ylab("Percent of Artifacts")+
  scale_y_continuous(labels = scales::percent) +
  geom_text(aes(label = paste0(round(n*100, digits = 0),"%")), position = position_stack(vjust = 0.5), s
  guides(fill=guide_legend(title="Rating"))+
  scale_fill_brewer(palette="Paired")+
  theme_light()
```



It appears that the artifacts were rated very similarly for this rubric. However, this similarity has skewed the association with the semester since more artifacts were from the fall semester. In other words, the similar scores means that the ratings are a reflection of the sample size for each semester. Therefore, it's possible that the inclusion of semester in the selection method model is biased from the disproportionate sample size and rating distribution.