TITLE:

PERFORMANCE IN GEN EDS: AN ANALYSIS ON STUDENT EVALUATIONS IN FRESHMAN STATISTICS

AUTHOR:
Anirban Chowdhury [achowdh1@andrew.cmu.edu]

ABSTRACT:

In this paper, we seek to understand how student ratings are distributed for the course Freshman Statistics in Dietrich College at Carnegie Mellon by examining the following: how the rating distributions differ across various rubrics, how much raters agree or disagree for each student or each rubric, how other factors like sex and semester affect ratings, and finally what other interesting properties exist in the data. We use data collected in an experiment from this statistics course in which 91 student assignments (called artifacts) were given to 3 separate raters and evaluated on a scale of 1-4 in 7 different categories. We used exploratory visualization to get a sense of the distributions of raters and rubrics, calculated rater disagreement by rubric, used multilevel models to assess correlations between rater scores and determine other important features in the data, and finally extended our visualizations and added diagnostics retrospectively to mine other important insights from the data. We found that raters tend to behave differently (one is harsher than the other two), and knowing the rater and the semester the course was taken allows us to gain important insights about a student's score (e.g. students in the spring tend to score lower across all rubrics as opposed to students in the fall) and come to the conclusion that our analysis suggests that there are meaningful differences in the behavior of the raters and in the way they treat each rubric, and that there are complex relationships between rater and rubric at play; in summary this means that the Dean's office should enforce more standardization on the experimental grading procedures in order to get a better sense of student performance.

INTRODUCTION:

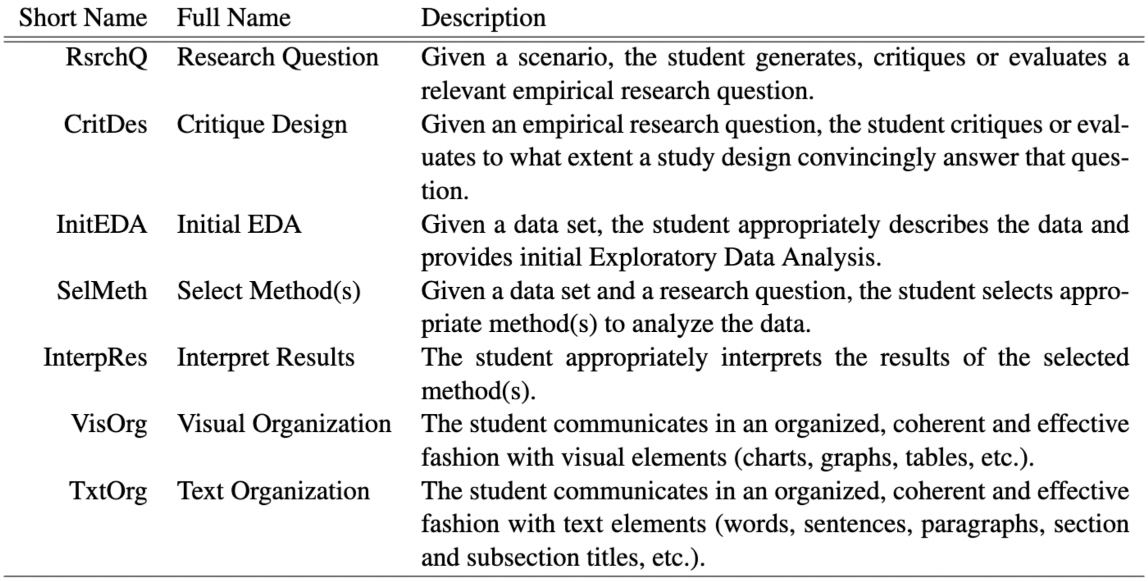
The Dietrich College of Humanities and Social Sciences at Carnegie Mellon provides its undergraduate students with a premier education in several foundational studies of social science. The college is particularly interested in student performance in the Freshman Statistics seminar, a course offered as a general education requirement to first-year students. In this work, we explore student performance as measured in 7 different rubric categories by 3 different raters from several angles. We address the following four research questions:

1. How do the distributions of various rubrics compare to one another? How do the distributions of rater's scores vary from one rater to another? Do certain rubrics or raters tend to be associated with higher or lower scores?
2. Within each rubric category, do raters generally agree on their scores? Is there any pattern in rater disagreement?
3. How do other factors, like student sex, semester the course was taken, etc. affect the ratings?
4. Are there any other interesting properties of the data that were not addressed in the previous three questions?

In this paper, we use a mixed effects regression analysis to study the relationships between student rating and these other features present in the data.

DATA:

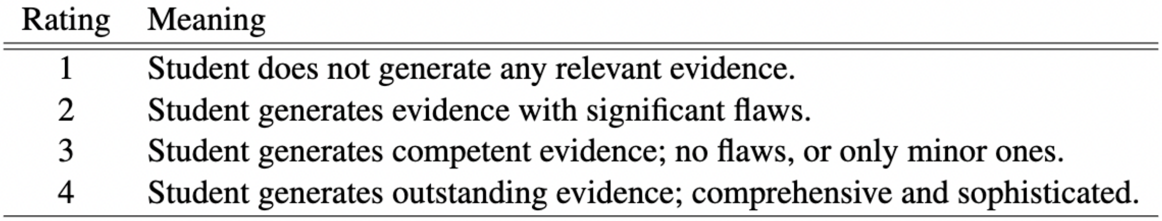
This research uses data collected from a Fall and Spring iteration of Freshman Statistics. 91 student assignments (called artifacts) were collected and sent to 3 raters for an evaluation in the following criteria in Figure 1:



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.).

FIGURE 1: Rubric Definitions

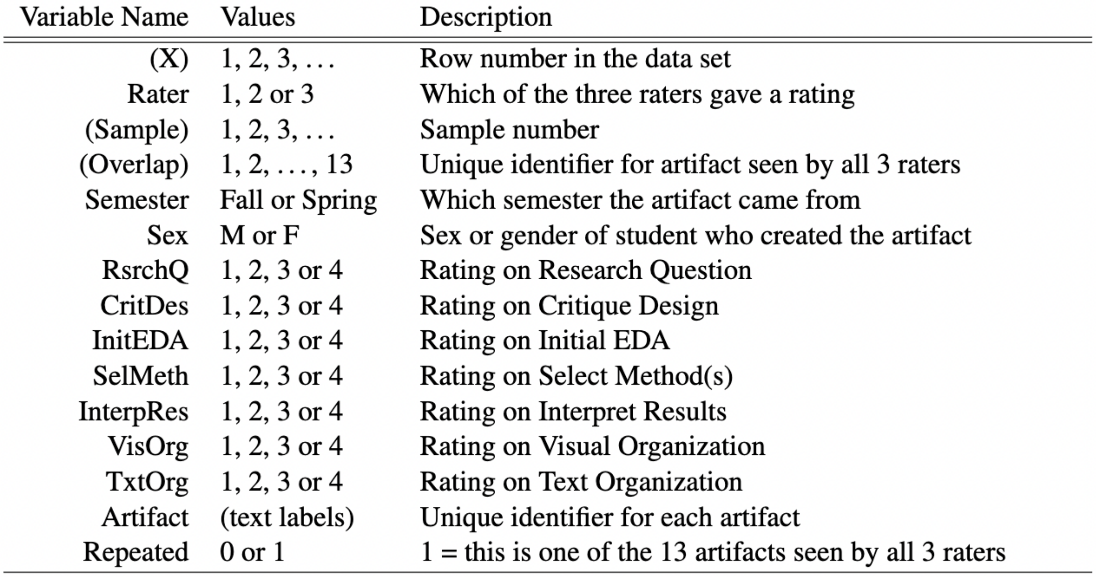
Each of these 7 rubrics was graded on a scale of 1 to 4 as follows in Figure 2:



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.

FIGURE 2: Rubric Grading Schemes

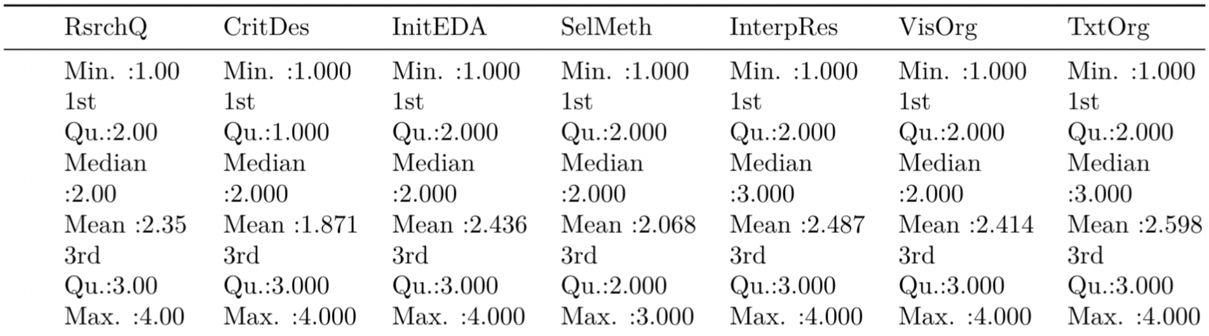
Note, this is an artificial grading scale conducted solely for this experiment. It does not reflect the actual grades earned by the students. 13 of the 91 artifacts were scored by all 3 raters, while the remaining 78 were only graded by one rater. So, we have 117 artifact/rater pairs in the dataset. For each artifact/rater pair, 7 scores from 1-4 were given, one for each rubric. In summary, we have the following features in our data as present in Figure 3:



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

FIGURE 3: Variables in the Dataset

There were a few data integrity issues. Most notably there was a missing value for Sex; we address this separately in the Methods section.



RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
Min. :1.00	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:2.00	1st Qu.:1.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000
Median :2.00	Median :2.000	Median :2.000	Median :2.000	Median :3.000	Median :2.000	Median :3.000
Mean :2.35	Mean :1.871	Mean :2.436	Mean :2.068	Mean :2.487	Mean :2.414	Mean :2.598
3rd Qu.:3.00	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000
Max. :4.00	Max. :4.000	Max. :4.000	Max. :3.000	Max. :4.000	Max. :4.000	Max. :4.000

FIGURE 4: Quantitative Variable Summaries

5-number summaries for the quantitative rubric variables in the dataset are presented in Figure 5. We can see that InterpRes and TxtOrg tend to get higher ratings (median score: 3) and SelMeth and CritDes have the lowest scores (by mean). For the categorical explanatory variables: each rater was assigned to exactly 39 artifacts, there were slightly more females than males (64 vs 52), and there were far more samples collected in the Fall than in the Spring (83 vs 34). Since most of the visualizations for this data are critical results for research question 1, we present them in the Results section. We also notice that there is a single missing value (encoded as a ”--”) for Sex in the dataset. We discuss the resolution in the Methods and Results sections.

This data is presented to us in two forms: a “wide” dataset with 7 rating columns, one for each rubric, and a “tall” dataset with one column for rating where every artifact/rater pair appears in 7 rows (one for each row) and an additional column for rubric type is added. We use both interchangeably as suits the analysis.

METHODS:

Data Preparation: Addressing Missing Values

Before conducting any analysis, we first use imputation to coerce a single missing value for the Sex column to one of the two classes. We examine histograms of each rubric cross-classified by Sex to determine how to best assign this value.

Research Question 1: How do the distributions of various rubrics compare to one another? How do the distributions of rater's scores vary from one rater to another? Do certain rubrics or raters tend to be associated with higher or lower scores?

To address this research question, we primarily focus on exploratory data analysis through statistical visualizations. In particular, we first examine histograms of ratings scores faceted by rater and compare the differences to determine if there are any qualitative variations in the rating distribution that these raters follow. Then, to assess whether or not these relationships change for rubric, we facet by rubric and add colored bars for each rater and again visually inspect the histograms for qualitative discussion.

Since not all raters graded every rubric, we repeat the above analysis using only the 13 artifacts that were seen by all 3 raters and determine if any meaningful differences in the variable relationships we discovered earlier can be found.

Research Question 2: Within each rubric category, do raters generally agree on their scores? Is there any pattern in rater disagreement?

To measure agreement between raters, we calculate the percent exact agreement for every possible pair of raters by counting up the number of times rater A and rater B agreed in their score for every artifact they both graded, for all possible A and B and dividing by the total number of ratings they gave out. That is, for every possible pair of 2 of the 3 raters, we calculate 7 quantities. For each rubric item for a set pair, we compute the sum of the number of times both raters gave a 1, the number of times both raters gave a 2, etc. and then divide by the number of ratings they gave out for that category. This quantity represents the proportion of times the raters agreed exactly on their ratings for a specific rubric. Then, we inspected these differences to see if there was any one rater that tended to lower agreement with the other two. By the nature of the problem, this analysis is only possible on the 13 artifacts that were seen by multiple raters.

To get a statistical measure of the correlation in rater scores, we fit 7 multilevel models regressing rating against a random effect with artifact as the grouping variable (one for each rubric), and measured the intraclass correlation (ICC) for these models. Because we are grouping by artifact, the observations in each group will correspond to the three raters. So, by measuring the ICC for these models, we can estimate the correlation between raters across all the artifacts. We repeat this step with various fixed effects as well. After computing all these ICCs, we inspect them for any apparent qualitative discoveries. We perform this procedure twice; once on all the artifacts in the dataset, and once with only the artifacts seen by all 3 raters in order to determine if there is any systematic difference in how artifacts were assigned to raters.

Research Question 3: How do other factors, like student sex, semester the course was taken, etc affect the ratings?

This research question focuses on a more general approach to understand how all features in the data relate to rating. We conduct this portion of the analysis in two steps.

First, we fit separate multilevel models for all 7 rubric ratings with an random intercept on artifact and add all fixed effects and multiple interaction terms for the variables Sex, Semester, Rater, and Repeated. For each of these models, we perform automated variable selection using BIC as our selection criterion to determine what variables and/or interaction terms help predict rubric rating, and then validate our results with manual inspection and ANOVA tests. We then try incorporating several possible random effects to determine if any would further improve the fit. We repeat this analysis both on the full dataset and then on the subset of artifacts seen by all 3 raters.

Since this approach does not let us directly assess interaction terms for rubric, we fit a final multilevel model to regress rating on data with rubric type as a feature using the full data to prevent any small sample size issues. We include sex, rubric, rater, and repeated as fixed effects, along with all interactions between rater, rubric, and semester as fixed effects, and we include a random slope for rubric grouped by artifact. We then perform automatic variable selection to determine which fixed effects help model ratings and try including several other random effects to assess their impact as well. This process again uses BIC as a selection criterion. We then inspect our reduced model and interpret our findings about what variables are important and assess their impact on rating.

Research Question 4: Are there any other interesting properties of the data that were not addressed in the previous three questions?

For this question, we perform supplementary EDA to examine the distributions of ratings as conditioned on other features like sex and semester to determine if there are any other important relationships besides those explicitly identified in our earlier EDA. In particular, we inspect the coefficients of our final multilevel model from Q3 and assess whether they agree with the qualitative observations we can make from the EDA, and if not, we offer possible explanations. We also consider model diagnostics and discuss implications of other possible models that could be better suited for this analysis to deal with issues in the diagnostics for the models we do present.

RESULTS:

Data preparation:

When examining the distributions of ratings across different rubrics and how they change across Sex, we can see that male and female students perform very similarly (see technical Appendix pages 2-18, relevant graph on page 10). Thus, with regards to the missing value, we simply impute it as Female (the mode of this column) because its value will likely be inconsequential to our analysis.

Research Question 1: How do the distributions of various rubrics compare to one another? How do the distributions of rater's scores vary from one rater to another? Do certain rubrics or raters tend to be associated with higher or lower scores?

As mentioned in the Methods section, we addressed this research question exclusively with visualizations.

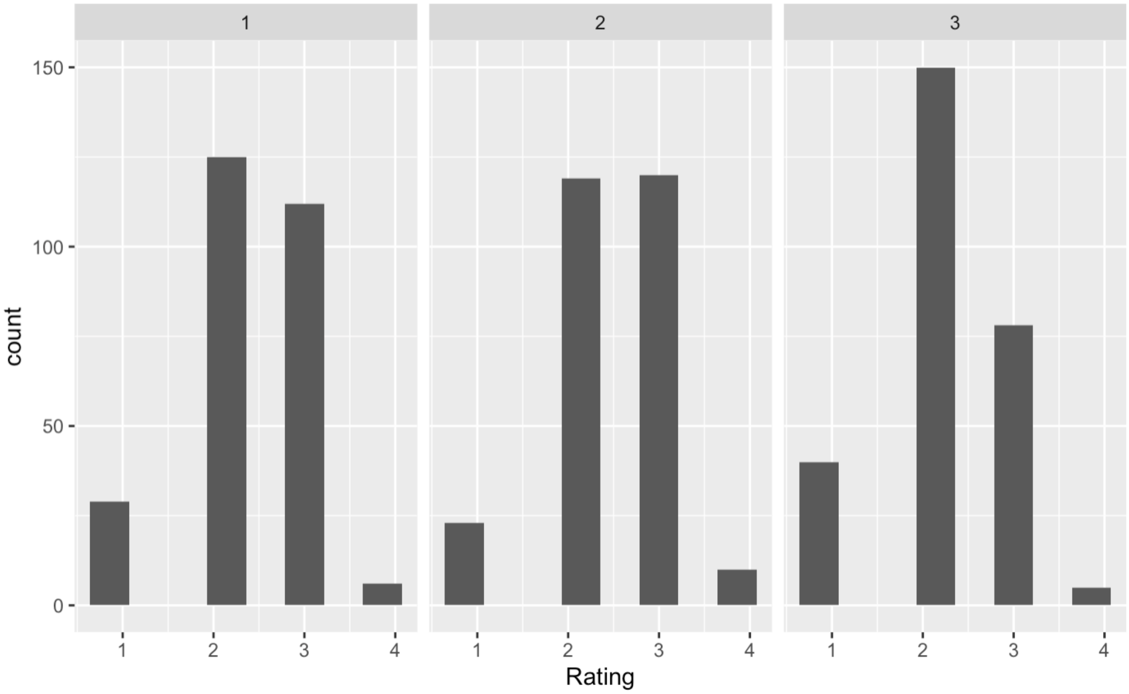


FIGURE 5: Rating Distributions by Rater

Our first visualization is presented in Figure 5. These three histograms represent the distribution of rating scores given out by the three raters. The most apparent observation we can make from these plots is that rater 3 seems to give out systematically lower scores (i.e. more 2's than 3's or 4's) while raters 1 and 2 seem to give out similar scores. This could be mirrored in the ICC analysis we perform later.

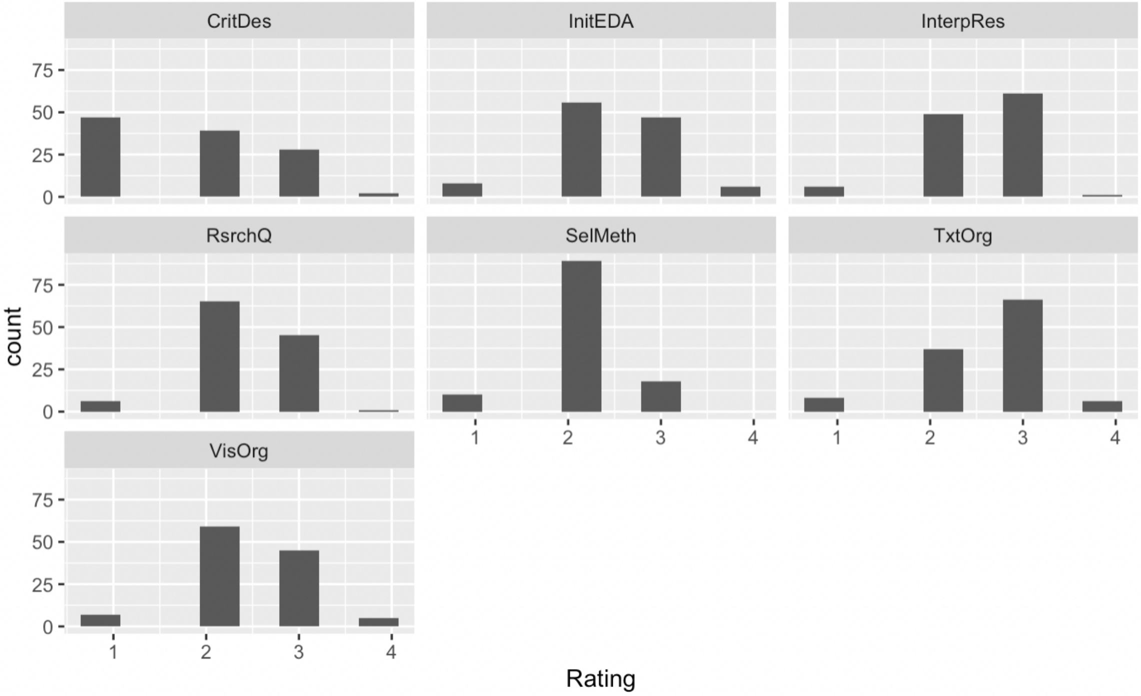


FIGURE 6: Rating by Rubric

Next, we examine how the ratings distributions differ across rubrics. The above faceted histograms represent rating counts for each of the 7 rubric categories, and we can immediately see several meaningful differences. For example, most of the scores for SelMeth are 2's, while InterpRes and TxtOrg are mostly scores of 3. There are also a very high proportion of 1's for CritDes, whereas there are almost no 1's for any other rubric item.

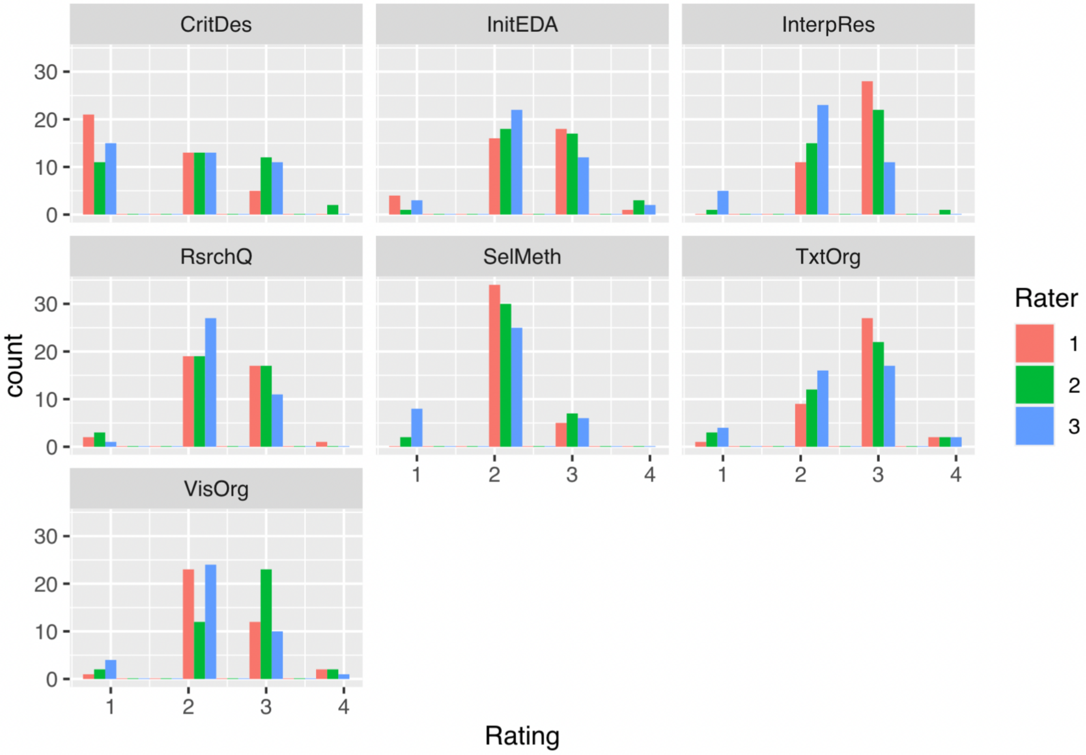
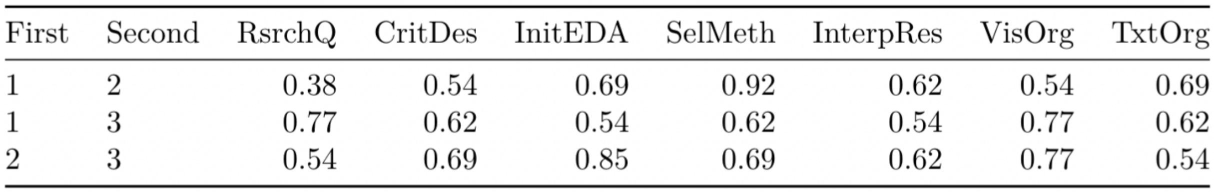


FIGURE 7: Rating by Rubric and Rater

Next, we combine these two analyses into one and present ratings histograms faceted on rubric category and colored by rater to determine if rater behavior differs across rubrics. There seem to be some slight differences in the rating distribution across raters for certain rubrics. For example, rater 1 gave out mostly 3s for InterpRes, but rater 3 gave out mostly 2s. However, for other rubrics like TxtOrg, the distribution seems similar. Looking at both these plots and the histogram aggregated over all rubrics, it seems in general like rater 3 was harsher (i.e. gave out lower scores) while rater 2 was more lenient (i.e. their scores seem to be generally higher), which matches our conclusions from the initial histograms. Overall, it seems like both rater and rubric are related to student ratings, as there are different rating distributions in their respective groupings, and there also appears to be some meaningful interaction between them, in that raters seem to behave differently across different rubrics. We performed these visualizations twice; the ones presented above were done on the full dataset. We also generated the same plots on only the artifacts that were observed by all 3 raters, but only a few meaningful differences were discovered (see technical appendix pages 2-18 for details). Because there is so little data for the common artifacts, it is difficult to say if the differences we observe are due to signal or small sample noise. For this reason, we proceed with the conclusions from the full data.

Research Question 2: Within each rubric category, do raters generally agree on their scores? Is there any pattern in rater disagreement?

For each of the possible pairs of 2 of the 3 raters, we calculate the percent exact agreement for every rubric. So, we have 21 entries in the following exact agreement table:

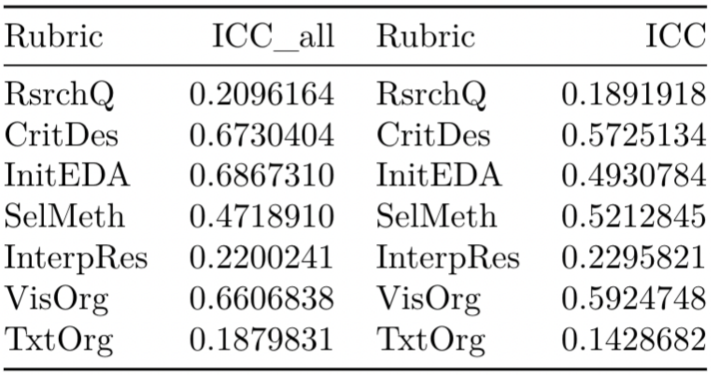


First	Second	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
1	2	0.38	0.54	0.69	0.92	0.62	0.54	0.69
1	3	0.77	0.62	0.54	0.62	0.54	0.77	0.62
2	3	0.54	0.69	0.85	0.69	0.62	0.77	0.54

TABLE 1: Percent Exact Agreement between Rater Pairs

The first observation that stands out is that there was 90% agreement between raters 1 and 2 for SelMeth, but only 20% agreement between these two for RsrchQ. This suggests a lot of variability across both raters and rubrics. In general it seems like raters 1 and 3 have slightly lower exact agreement quantities (most are around 0.5-0.6), which makes sense given our observations from the EDA. Rater 2 seems to have a marginally better agreement with both raters, which makes sense. One important characteristic is that there is variation across rater pairs, i.e. high agreement in one rubric item for a certain pair does not seem to indicate that other pairs will necessarily also agree highly for the same item.

In order to get a sense of the correlation between all three raters for each rubric item, we can look at the intraclass-correlation scores from multilevel modeling. When Artifact is a grouping variable, all observations per group will correspond to the raters' scores for that artifact. Thus, the ICC will measure the correlation between the three raters across all the artifacts. We can calculate these quantities by fitting random intercept models for each rubric response. We fit 14 of these models in total: 7 for the full dataset, and 7 for the data that was seen by every rater in order to inspect for any differences in how artifacts were assigned, and calculate the ICC for each one and present the results in the table below.



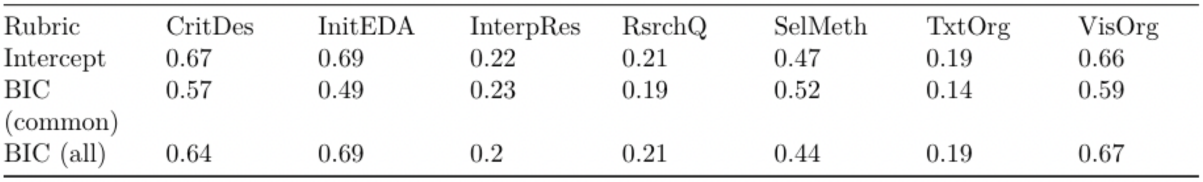
Rubric	ICC_all	Rubric	ICC
RsrchQ	0.2096164	RsrchQ	0.1891918
CritDes	0.6730404	CritDes	0.5725134
InitEDA	0.6867310	InitEDA	0.4930784
SelMeth	0.4718910	SelMeth	0.5212845
InterpRes	0.2200241	InterpRes	0.2295821
VisOrg	0.6606838	VisOrg	0.5924748
TxtOrg	0.1879831	TxtOrg	0.1428682

TABLE 2: ICC for Rubrics across All Data and Repeated Data

The table above presents the ICCs for every rubric. ICC_all denotes the score for all 91 artifacts, and ICC denotes the score for the 13 repeated artifacts. The first finding here is that the ICCs are fairly similar across the full and repeated data, the only large difference is InitEDA. These scores show us that for Research Question, there is very little agreement among raters, while for Visual Organization there is higher agreement across raters. This could mean that the raters have different perceptions regarding how they grade the artifacts, i.e. one rater might think certain characteristics lead to a good research question while another rater thinks different characteristics might lead to a good research question. Alternatively, this could mean that all the raters have similar ideas of what makes good visual organization, as the ICC is higher for this rubric iterm. To view the fitted models and their estimated coefficients, see the technical appendix pages 18-27.

Research Question 3: How do other factors, like student sex, semester the course was taken, etc. affect the ratings?

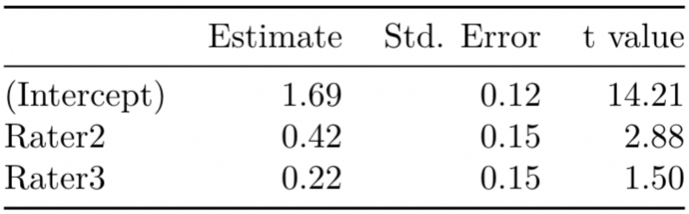
We now extend our multilevel models above to include fixed effects in order to capture other variables in the model that could have meaningful relationships with an artifact's rating. For each of the 14 models presented above, we add fixed effects for every variable we have in the data and then do variable selection to determine which fixed effects could improve the fit in terms of BIC. So, in total for each rubric, we have the following models: a random intercept model on the full data, a random intercept model on the common data, and a variable selected model with the highest BIC, using the full data. We also tried variable selection on the common data, but no other features besides the random intercept improved the fit, so we omit it from interpretation.



Rubric	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg
Intercept	0.67	0.69	0.22	0.21	0.47	0.19	0.66
BIC (common)	0.57	0.49	0.23	0.19	0.52	0.14	0.59
BIC (all)	0.64	0.69	0.2	0.21	0.44	0.19	0.67

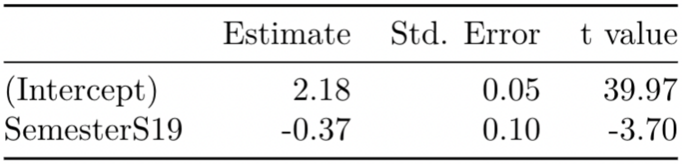
TABLE 3: ICC for Rubrics with best BIC Models

The first row in Table 3 represents rubric type, the second row is the ICC for the intercept model with an artifact grouping variable, the third row is the ICC for the highest BIC model we found for each rubric with the same random effect using only the 13 common artifacts, and the last row is the highest BIC model for all the data. The coefficients and random effects for all of these models are present in the technical appendix, pages 29-53. Most of these ICCs are very similar across all rubrics, suggesting that in most cases adding fixed effects for rater, semester, sex, repeated, etc. does not impact agreement between raters. However, the ICCs for SelMeth, VisOrg, and InitEda are in fact different for the variable selected models than the intercept one. This suggests that our variable selection for semester actually does in fact influence rater agreement, or that these rater's scores tend to be more correlated when adjusting for fixed semester effects. This suggests two things: The semester and rater terms could be meaningful in determining ratings, and that the relationship between these terms differs across rubric.



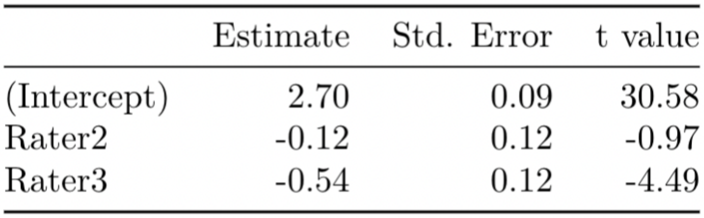
	Estimate	Std. Error	t value
(Intercept)	1.69	0.12	14.21
Rater2	0.42	0.15	2.88
Rater3	0.22	0.15	1.50

CritDes



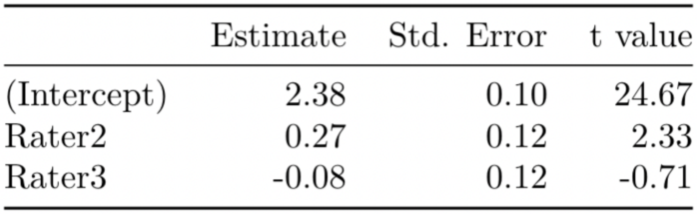
	Estimate	Std. Error	t value
(Intercept)	2.18	0.05	39.97
SemesterS19	-0.37	0.10	-3.70

SelMeth



	Estimate	Std. Error	t value
(Intercept)	2.70	0.09	30.58
Rater2	-0.12	0.12	-0.97
Rater3	-0.54	0.12	-4.49

InterpRes



	Estimate	Std. Error	t value
(Intercept)	2.38	0.10	24.67
Rater2	0.27	0.12	2.33
Rater3	-0.08	0.12	-0.71

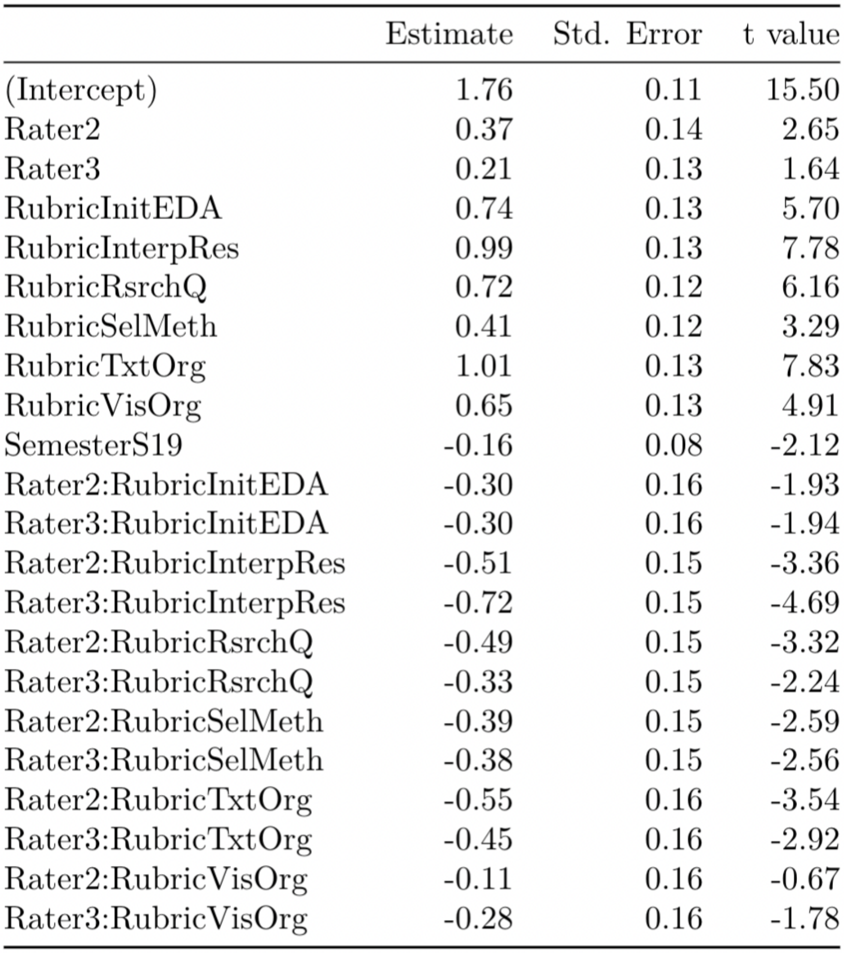
VisOrg

TABLE 4: Fixed Effects for each Non-Intercept Rubric Model

When inspecting these models in detail, we can find specific variables that influence rating in certain rubric categories. For InitEda, TxtOrg, and RsrchQ, the highest BIC model we found, both by manual inspection and automatic selection, was from the random intercept model. For SelMeth, the highest BIC included a fixed effect for Semester, and for InterpRes, CritDes, and VisOrg, the highest BIC model included Rater as a fixed effect. Again, this suggests that the relationship between rater and rating changes across different rubrics, and the relationship between semester and rating also changes across different rubrics. The tables in Table 4 display the fixed effect coefficients for CritDes, SelMeth, InterpRes, and VisOrg, i.e. all the rubrics where variable selection led to a model with more than just a random intercept. By examining these coefficients, we can make conclusions about the specific impact these variables have on rating for each rubric. For the CritDes rubric, we can see that students would expect a 0.2 unit increase in their score if they were to be graded by rater 2 as opposed to rater 3, and another 0.2 unit increase were they to be graded by rater 3 as opposed to rater 1, controlling for the random intercept. Thus, for this rubric rater 1 is the harshest grader, which matches the EDA from page 9. For the SelMeth rubric, students in the spring are expected to perform 0.37 units worse on average than students in the fall, indicating either rater perception of what qualifies as a good SelMeth category varied from semester to semester or the course teaching varied more from the rubric definition of SelMeth in the spring. For InterpRes, students would expect a 0.5 unit decrease in score on average were they to be graded by rater 3 as opposed to rater 1, or a 0.1 unit decrease were they to be graded by rater 3 as opposed to rater 2. Thus, rater 3 is the harshest for this rubric item. Similarly, for VisOrg, students would expect about a 0.1 decrease in score were they to be graded by rater 3 as opposed to rater 1, and about a 0.35 unit decrease were they to be graded by rater 3 instead of rater 2. In summary, these models quantitatively tell us that raters behave differently per rubric, i.e. certain raters are harsher on certain rubrics and more lenient on others.

Another interesting finding in this modeling step relates to the random effects. For each model, we can treat the random effect standard error as a measure of the variation in the rating that cannot be captured by the fixed intercept. When inspecting these values closely (see technical appendix pages 36-53) we find that CritDes, InitEDA, and VisOrg all have much higher variances than the other rubrics (0.3-0.4 as opposed to 0.06-0.08). This could indicate a possible lack of common understanding of rubric definitions among the raters. If raters tend to give out different scores for these rubrics, it could mean that each rater has different qualifications of what makes an artifact score well in these categories, or that the grading criteria for these rubric items are unclear to the raters.

The final step in our modeling process is now to add rubric as a variable in our model in order to assess its interactions with the other variables we have more directly. To this end, we swap to the “tall” data and fit a model with rubric as a fixed effect and explore its interactions with the other variables. We start with a full model using fixed effects for sex and repeated, as well as fixed effects and interactions with rater, rubric, and semester. We experiment with BIC-based automatic variable selection to delete unnecessary fixed effects and introduce new random effects to arrive at the fixed effects in the final model presented in Table 5.



	Estimate	Std. Error	t value
(Intercept)	1.76	0.11	15.50
Rater2	0.37	0.14	2.65
Rater3	0.21	0.13	1.64
RubricInitEDA	0.74	0.13	5.70
RubricInterpRes	0.99	0.13	7.78
RubricRsrchQ	0.72	0.12	6.16
RubricSelMeth	0.41	0.12	3.29
RubricTxtOrg	1.01	0.13	7.83
RubricVisOrg	0.65	0.13	4.91
SemesterS19	-0.16	0.08	-2.12
Rater2:RubricInitEDA	-0.30	0.16	-1.93
Rater3:RubricInitEDA	-0.30	0.16	-1.94
Rater2:RubricInterpRes	-0.51	0.15	-3.36
Rater3:RubricInterpRes	-0.72	0.15	-4.69
Rater2:RubricRsrchQ	-0.49	0.15	-3.32
Rater3:RubricRsrchQ	-0.33	0.15	-2.24
Rater2:RubricSelMeth	-0.39	0.15	-2.59
Rater3:RubricSelMeth	-0.38	0.15	-2.56
Rater2:RubricTxtOrg	-0.55	0.16	-3.54
Rater3:RubricTxtOrg	-0.45	0.16	-2.92
Rater2:RubricVisOrg	-0.11	0.16	-0.67
Rater3:RubricVisOrg	-0.28	0.16	-1.78

TABLE 5: Fixed Effects for Combined Model

So, fixed effects for rater, semester, rubric, as well as interactions between rater and rubric were all necessary in improving the BIC to predict ratings. From these results we can gather that rubric, rater, and semester are all meaningful predictors of rating, and that the relationship between rater and rating changes across different rubrics. For example, we expect the increase in score from rater 2 to rater 3 to be lower for vis org rubrics than other rubrics when controlling for other variables in the model. We can also say that the relationship between rubric and ratings, and the relationship between rater and ratings are different across groups, hence the importance of the random effects, i.e. for each artifact we can expect different ratings across raters for different rubrics.

When examining the coefficients of the model, we can make conclusions about the specific impact of all these fixed effects on rating. For example, the coefficient for SemesterS19 is negative, meaning that when controlling for all other variables, we expect students taking the course in the Spring to have a lower rating on average than those taking it in the Fall. Additionally, when holding all other variables constant, the model expects the rating for CritDes to be lower than all other ratings, which matches our findings from the EDA section. The coefficient for rater 3 is positive, suggesting at first glance that this rater gives out higher scores (contradicting our EDA). We will examine this more closely in Research Question 4.

We can also make interesting discoveries about the data when examining the random effects of our complete model (see technical appendix pages 53-63). We can examine the estimated variance of all of the random slopes of rubric grouped by artifact and notice that there is much less variation in ratings for Selection Method across all the artifacts than the other items, and fairly more variation in CritDes. This variation cannot be captured by the fixed effects and shows us that there are inherent intricacies in the data that should be investigated closely to draw inferences. It shows us that the random and fixed effects, as well as the interactions, are meaningful, and should be taken into consideration with the conclusions drawn from earlier visualizations.

In summary, our findings from these models indicate several important features of the dataset. First, there are differences in what variables affect rating and which do not when examining all the data versus just the repeated data. However, this is likely due to the repeated data being very small in size rather than a meaningful statistical signal. Second, the variables that impact rating change from rubric to rubric, meaning the relationships between some of these variables (namely rater and semester) and rubric vary depending on the type of rubric. Finally, there are meaningful interactions between raters and rubrics, meaning that the three raters seem to behave differently across the different rubrics, and the degree of this variation changes from rubric to rubric.

Research Question 4: Are there any other interesting properties of the data that were not addressed in the previous three questions?

Throughout this work we focused specifically on how raters and rubrics interacted and influenced rating, so in this section we use more EDA to explore how the relationships between Sex, Semester, and Repeated influence rating. As discussed in the technical appendix (pages 2-18), we do not see any visual differences in the rating distributions across student sex, and this conclusion was mirrored in the modeling procedures where Sex never improved the fit for any rubric.

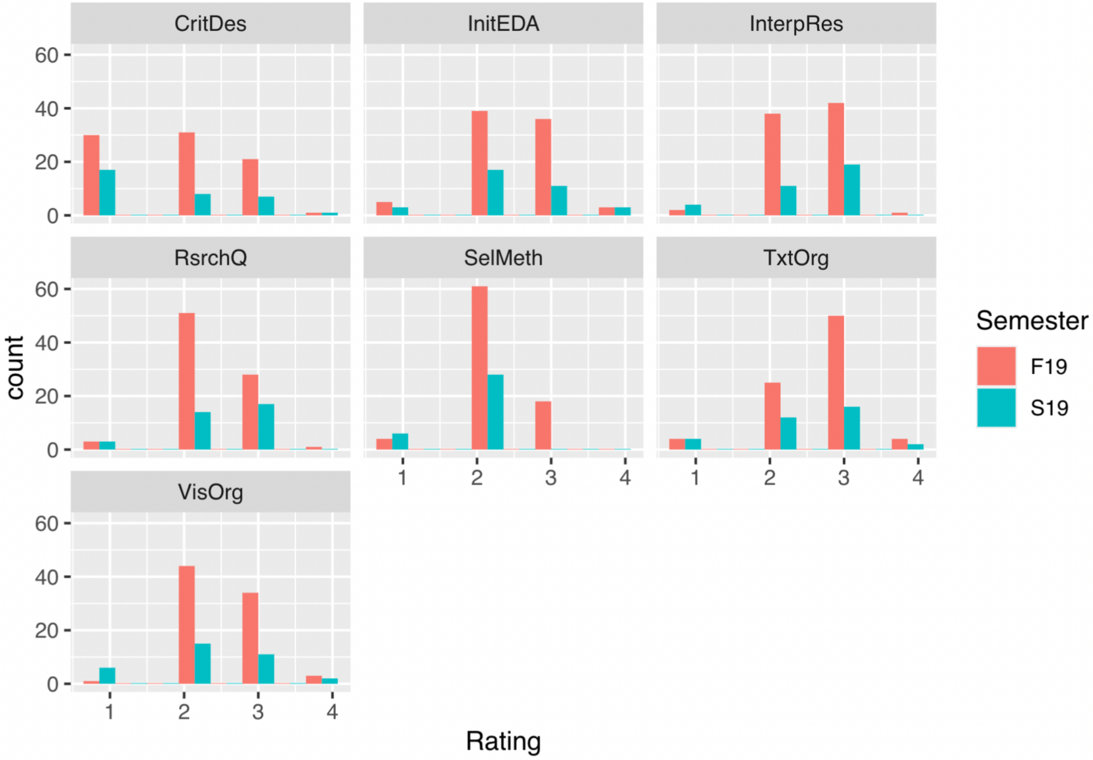


FIGURE 8: Rating by Rubric and Semester

Above is a similar faceted histogram where we split the data based on whether or not the artifact was from the Fall or Spring. We can see a few key differences in rating distribution from these two groups. For SelMeth, no 3s were given in spring but around 20 were given in the fall. Also, for RsrchQ, the most common rating in fall was a 2 but it was a 3 in the spring. If we repeat this analysis for variables like Sex or repeated, we will not be able to observe any key differences (see technical appendix pages 2-18).

Next, we closely examine the coefficients of the fixed effects and interactions of our final model and determine if they match up with our expectation from the EDA we did in Question 1. There is an apparent discrepancy between our EDA findings and our final multilevel model, in that although we determined from EDA that rater 3 was harsher and tended to give out lower scores, the model estimate for the increase in rating associated with rater 3 as opposed to rater 1 was positive, indicating that rater 3 actually gives out higher scores. However, this is only the case when controlling for all the other fixed effects and interactions that were not present in the EDA. When accounting for interactions, the model expects rater 3 to give out lower scores on average for most rubrics, which matches our findings for Research Question 1. This suggests again that rater behavior cannot be captured without considering an artifact's rubric category.

Finally, we examine model residuals to determine if a linear model is an appropriate fit. Because the response is discrete with four levels, it is possible that a multilevel multinomial logistic regression model would be more appropriate. Thus, we examine a plot of the Cholesky residuals against fitted values as present in the technical appendix page 59-63. It seems like there is some nonconstant variance, although this is more likely to be a result of varying sample size. We see less variation in the errors at smaller fitted values, but we also expect there to be few samples with very low scores, so this low variation makes sense. Overall, this diagnostic plot seems reasonable and it appears that a linear model is indeed appropriate, at least for this specific assumption of constant variance.

DISCUSSION:

In this paper, we present a thorough statistical analysis on the experimental data provided from Dietrich College regarding student performance in Freshman Statistics in order to evaluate student performance, rater consistency, and any other meaningful information for the Dean's Office. We specifically focused on determining how ratings change across rubrics or raters, how raters agree or disagree, what other variables are useful in predicting rating, and any other miscellaneous insights that can be found. Based on our findings, our largest actionable takeaways for the Dean's Office are that raters tend to behave differently across different rubrics and behave differently from each other and that both rater and semester are useful variables when modeling student performance. In context, this means that it is difficult to directly assess or compare student performance due to these intricate relationships with raters, rubrics, and semester. In order for Dietrich College to get a fair measurement of student learning, more standardization of rater behavior and rubric definitions should be imposed across all iterations of the course to ensure an accurate representation of performance and understanding.

We used EDA through data visualization to address the first question. By visually inspecting faceted histograms, we found that the raters exhibit different behavior, and that their behavior changes with the rubric category. We calculated percent exact agreement and ICC for random intercept models for every rubric in order to get quantitative metrics on how much the raters tend to agree or disagree, and found that one rater tends to disagree more than others and patterns of agreement vary across rubrics. To address the third question, we fit multilevel models to regress rating against multiple factors, considering both fixed and random effects, and we discovered that rater, semester, rubric, and the interaction between rater and rubric are all important in modeling rating. For the fourth question, we explored some additional EDA and determined that the other factors we did not explicitly examine to answer Question 1 (like Sex) are not meaningful predictors or rating, while others (like Semester) are. We also discovered that The distribution of ratings and relationships with the other factors does not change much when considering the data with only 13 artifacts that all raters saw versus the data with all the rubrics.

In context, this work allows us to present stakeholders in Dietrich College with several specific findings. First, the raters did not all exhibit similar behavior, so students who received poorer scores from rater 3 might actually have done well in the eyes of the other two. Also, semester seems to be an important predictor of rater, and in particular we estimate that students that took the course in the spring would do slightly worse than those who took it in the fall (controlling for other factors). So, the experimental grading of the course was not consistent across these semesters. Furthermore, male and female students performed equally well.

A clear extension of this work is to use a more appropriate modeling paradigm, like multilevel multinomial logistic regression. Rating is likely more appropriately treated as a categorical variable since it is discrete with only four levels. The diagnostic plots we examined for a linear fit seemed reasonable, but it is possible that the small differences in variation we saw were due to an inappropriate modeling setup. The plot we examined also might not have been extensive in determining all possible issues with the fit. Thus, reattempting modeling with a generalized linear mixed model is a natural extension of our work. Additionally, we were relatively constrained in terms of the variables we attempted to include in our models. Although several fixed and random effects were explored, some were not able to be fit due to the small sample size. So, repetition of this experiment in different years to expand group sizes would allow us to examine more possible intricacies in the data.

BIBLIOGRAPHY:

Junker, B. W. (2021). *Project 02 assignment sheet and data for 36-617: Applied Regression Analysis.* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh PA. Accessed Nov 29, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

Junker, B. W. (2021). *HW10 Solutions for 36-617: Applied Regression Analysis.* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh PA. Accessed Nov 29, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

Sheather, Simon J. (2010). *A Modern Approach to Regression with R*. New York: Springer.

36-669 Project 02 Technical Appendix

Anirban Chowdhury

11/13/2021

Contents

```
source('residual-functions.R')
library(arm)
```

```
## Loading required package: MASS

## Loading required package: Matrix

## Loading required package: lme4

##
## arm (Version 1.12-2, built: 2021-10-15)

## Working directory is /Users/anirbanchowdhury/Downloads
```

```
library(lme4)
library(ggplot2)
library(plyr)
```

Note: The text in this appendix highlights the same points expressed in the paper (in more detail) to that in the paper to allow for clear reading. The appendix follows the same structure as the paper (split into sections for each question) with text describing main points of the code that mirrors the important findings highlighted in the paper.

Research Question 1 (Complete EDA)

How do the distributions of various rubrics compare to one another? How do the distributions of rater's scores vary from one rater to another? Do certain rubrics or raters tend to be associated with higher or lower scores?

To determine the relationships between the variables in the ratings dataset, with particular attention to ratings and rubrics, we do some EDA. We first generate summary statistics of the quantitative rubric scores, and count tables of rater, sex, and semester. Note that there were some missing values for sex present in the data, so we impute them with female (the mode). As we will show below, the distribution of ratings across rubrics does not vary much with gender, so how we assign this sex will not really impact our analysis.

```
library(knitr)
ratings = read.table('ratings.csv', sep = ",", header = T)
ratings[5, 'Sex'] = "F"
tall = read.table('tall.csv', sep = ",", header = T)
tall$Rater = as.factor(tall$Rater)
ratings$Rater = as.factor(ratings$Rater)
tall[which(tall$Sex == ""),"Sex"] = 'F'
kable(summary(ratings[,c(7:13)]))
```

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
	Min. :1.00	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
	1st Qu.:2.00	1st Qu.:1.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000
	Median :2.00	Median :2.000	Median :2.000	Median :2.000	Median :3.000	Median :2.000	Median :3.000
	Mean :2.35	Mean :1.871	Mean :2.436	Mean :2.068	Mean :2.487	Mean :2.414	Mean :2.598
	3rd Qu.:3.00	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000
	Max. :4.00	Max. :4.000	Max. :4.000	Max. :3.000	Max. :4.000	Max. :4.000	Max. :4.000
	NA	NA's :1	NA	NA	NA	NA's :1	NA

```
table(ratings$Rater)
```

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

```
table(ratings$Sex)
```

```
##
##  F  M
## 65 52
```

```
table(ratings$Semester)
```

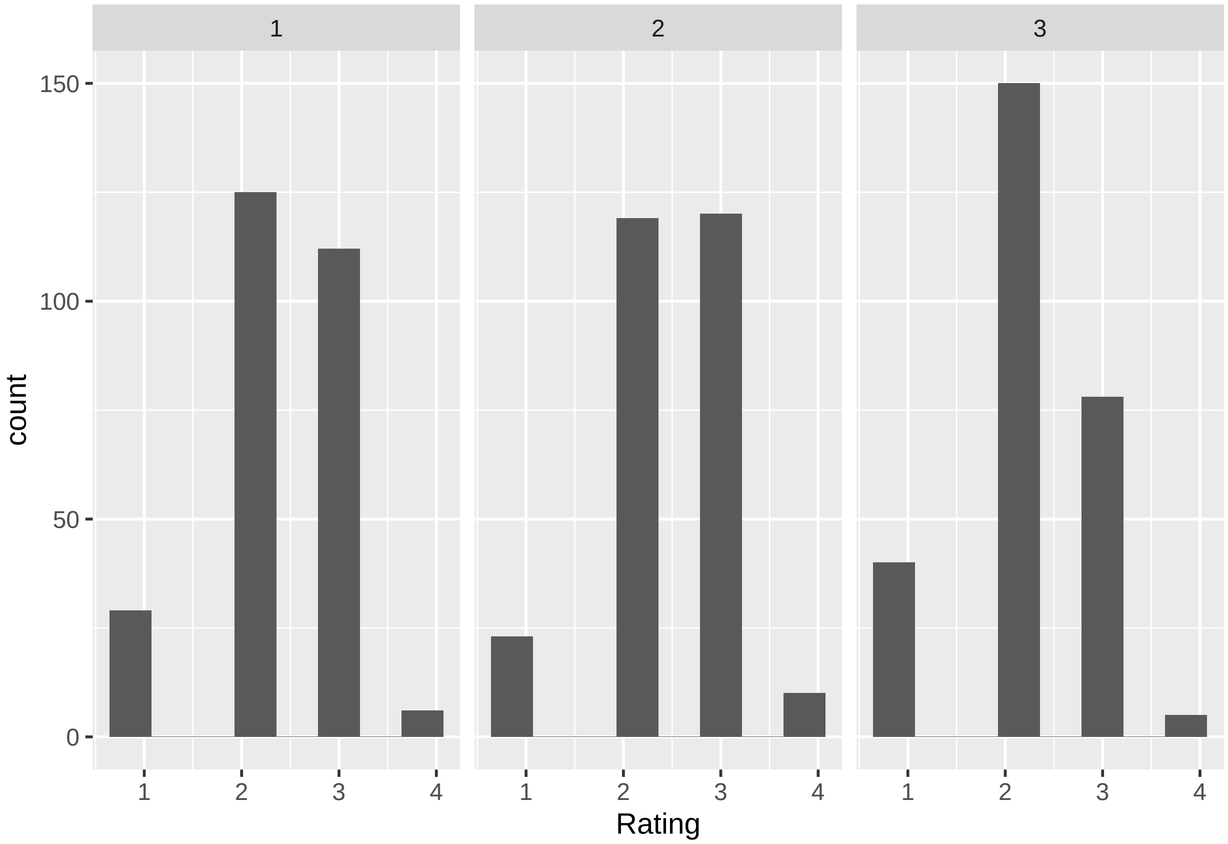
```
##
##   Fall Spring
##     83     34
```

From count tables, we can see that there are more females than males in the data, and there are equal samples across rater. There are also more samples from the fall semester than the spring.

It seems that InterpRes and TxtOrg have higher scores in general than the other rubric items. We can look at histograms to confirm this.

```
ggplot(data = tall ) + geom_histogram(aes(x = Rating), bins = 8,
                                      position = 'dodge') + facet_wrap(~Rater)
```

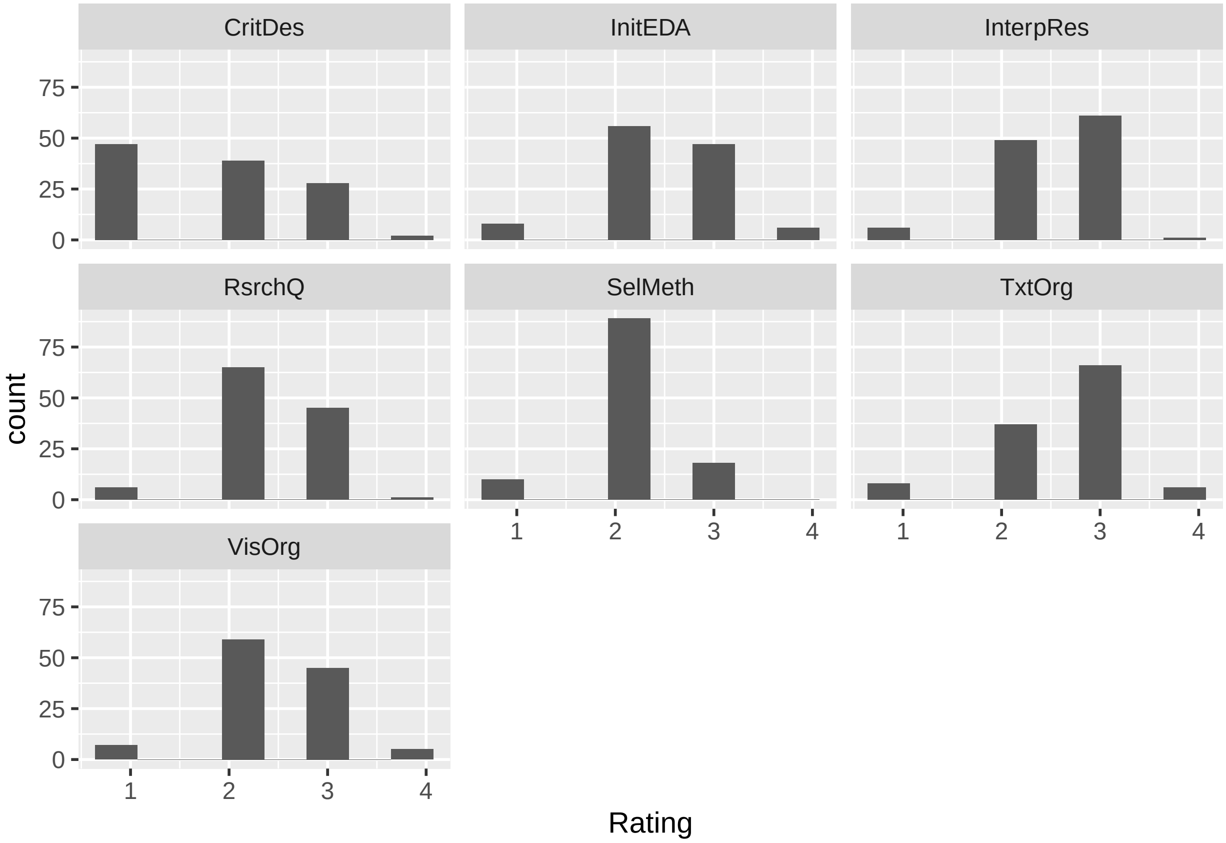
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



We see different rating patterns across the three raters, and notice especially that rater 3 tends to give lower scores.

```
ggplot(data = tall) + geom_histogram(aes(x = Rating), bins = 8,
                                     position = 'dodge') + facet_wrap(~Rubric)
```

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

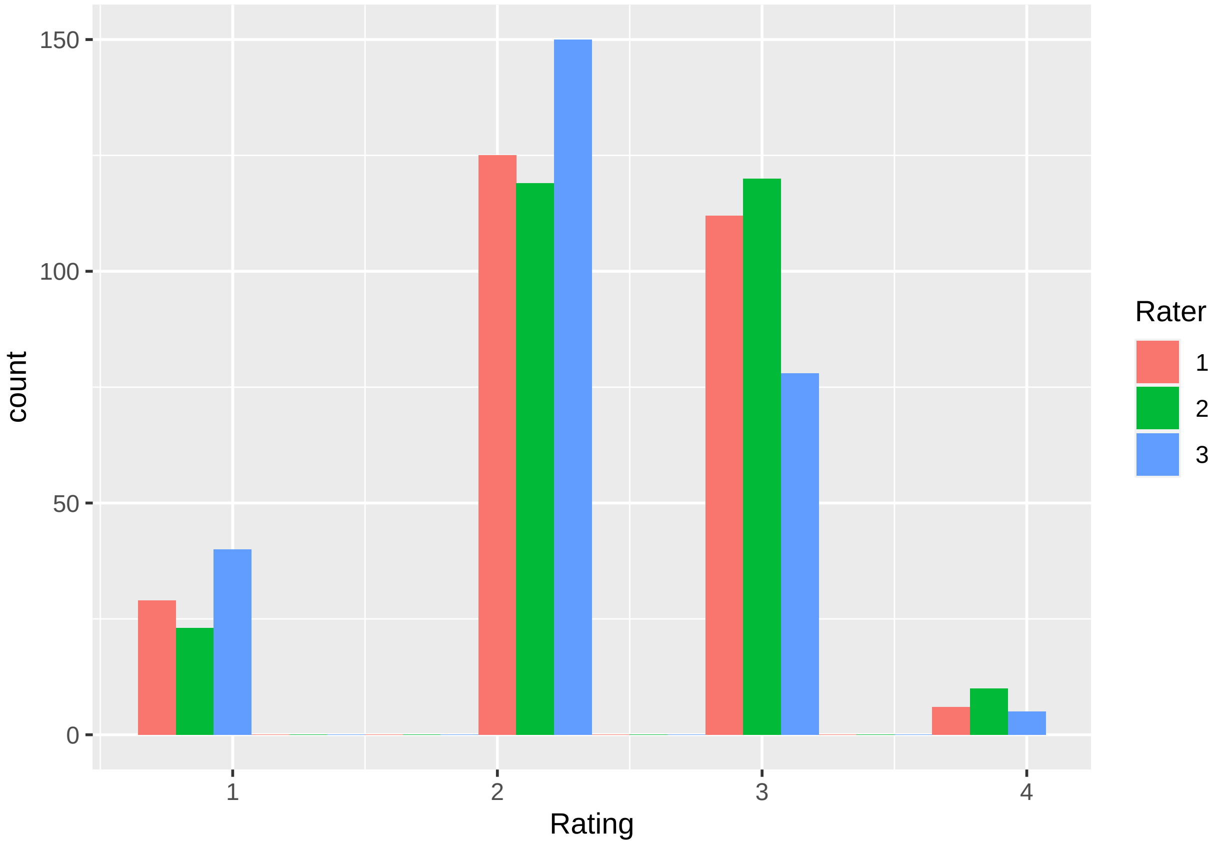


these histograms show us that the rating distribution is different across rubrics, e.g. text org and interp res ratings tend to be distributed higher and skewed left when compared to sel meth.

Next, we examine how rubric rating distributions change over certain factors in the dataset.

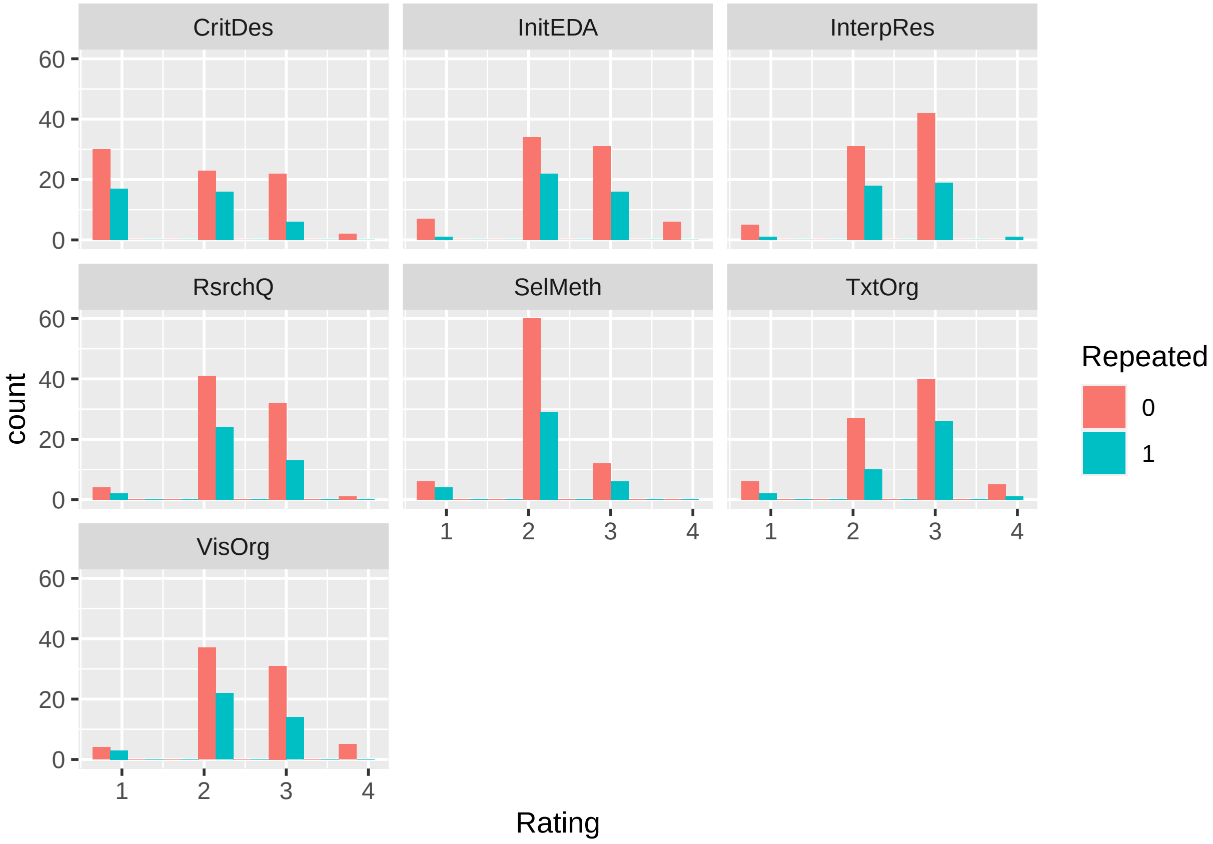
```
ggplot(data = tall) + geom_histogram(aes(x = Rating, fill = as.factor(Rater)),
                                     bins = 8, position = 'dodge') +
  guides(fill=guide_legend(title="Rater"))
```

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



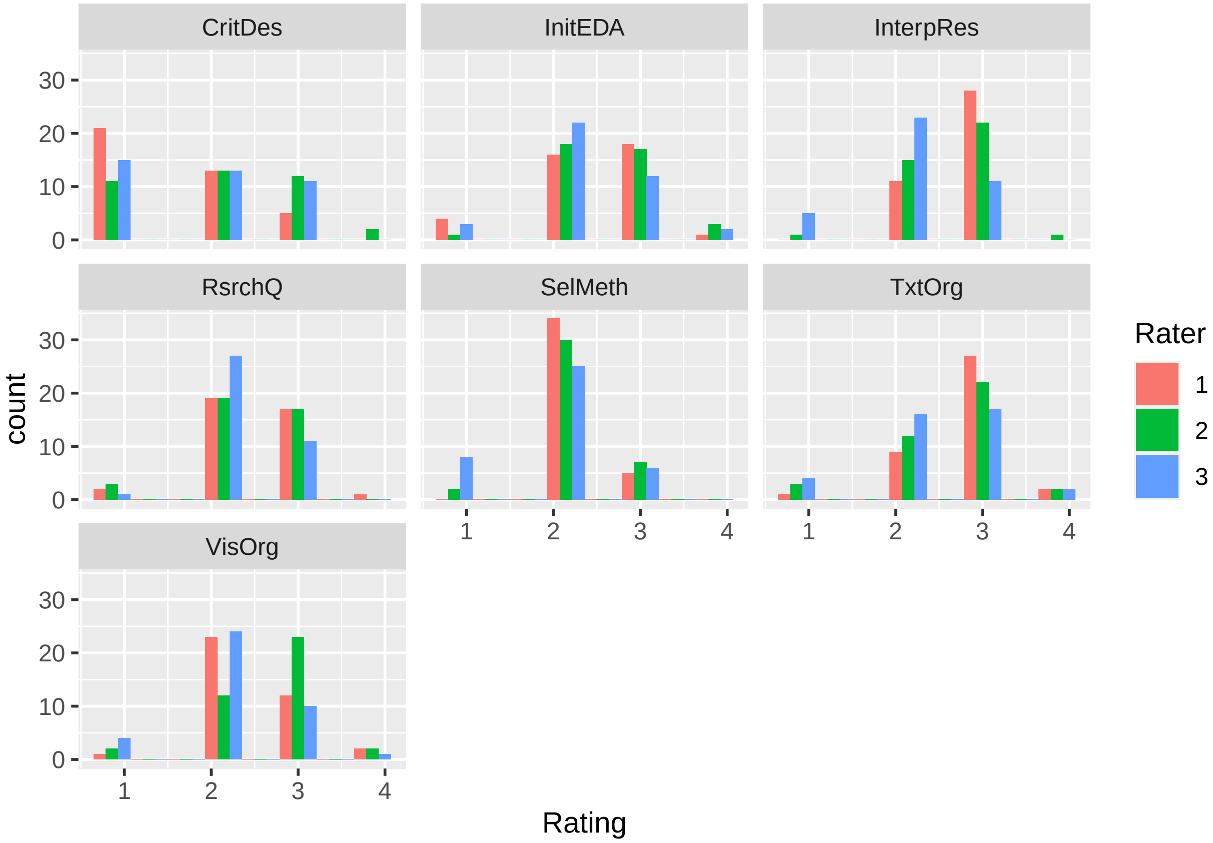
```
ggplot(data = tall) + geom_histogram(aes(x = Rating,
                                         fill = as.factor(Repeated)),
                                     bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Repeated"))
```

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



```
ggplot(data = tall) + geom_histogram(aes(x = Rating, fill = as.factor(Rater)),
                                     bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Rater"))
```

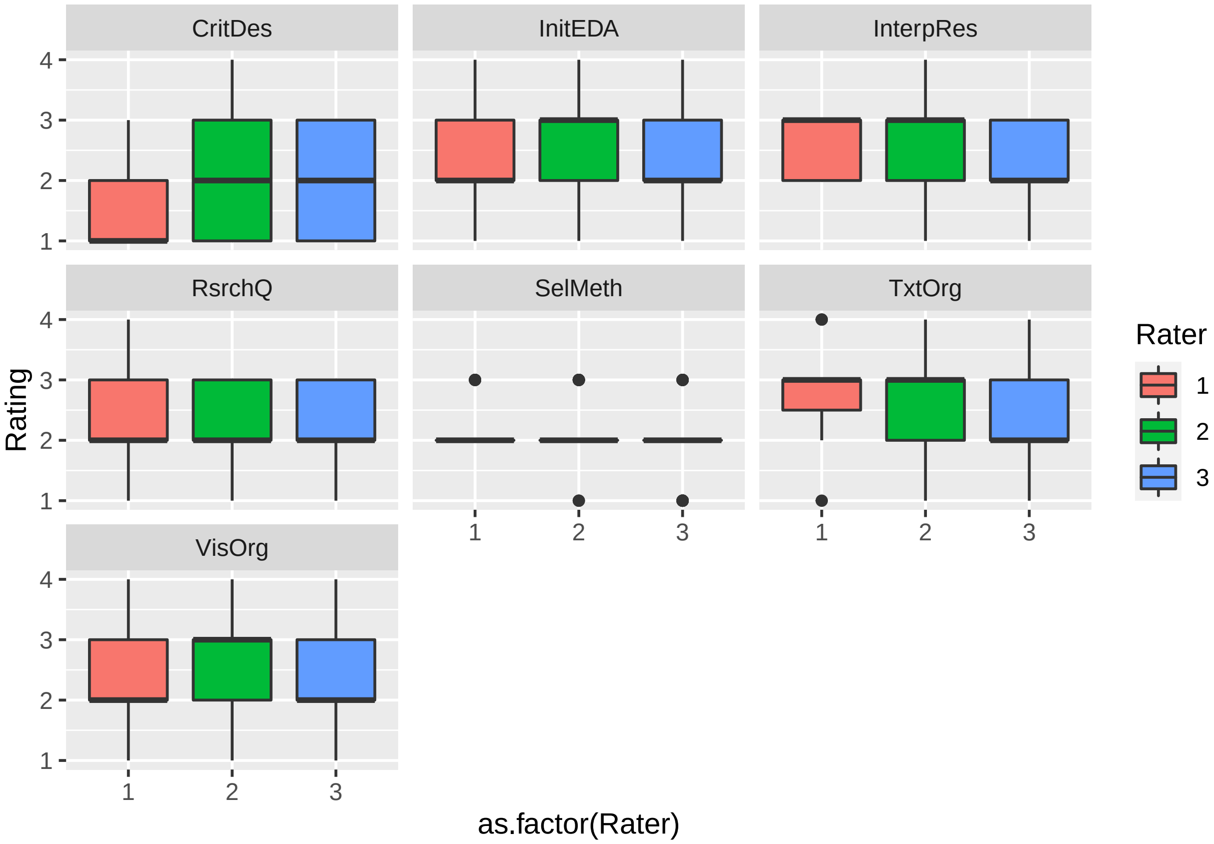
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



There are a few differences in the histograms above. In general rater 3 seems to give out lower scores and rater one seems to give out higher scores. However, this is dependent on rubric, i.e. for crit des rater 1 gave out mostly 1s. So, rater behavior seems to differ both across raters and across rubrics.

```
ggplot(data = tall) + geom_boxplot(aes(y = Rating, x = as.factor(Rater),
                                       fill = as.factor(Rater))) +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Rater"))
```

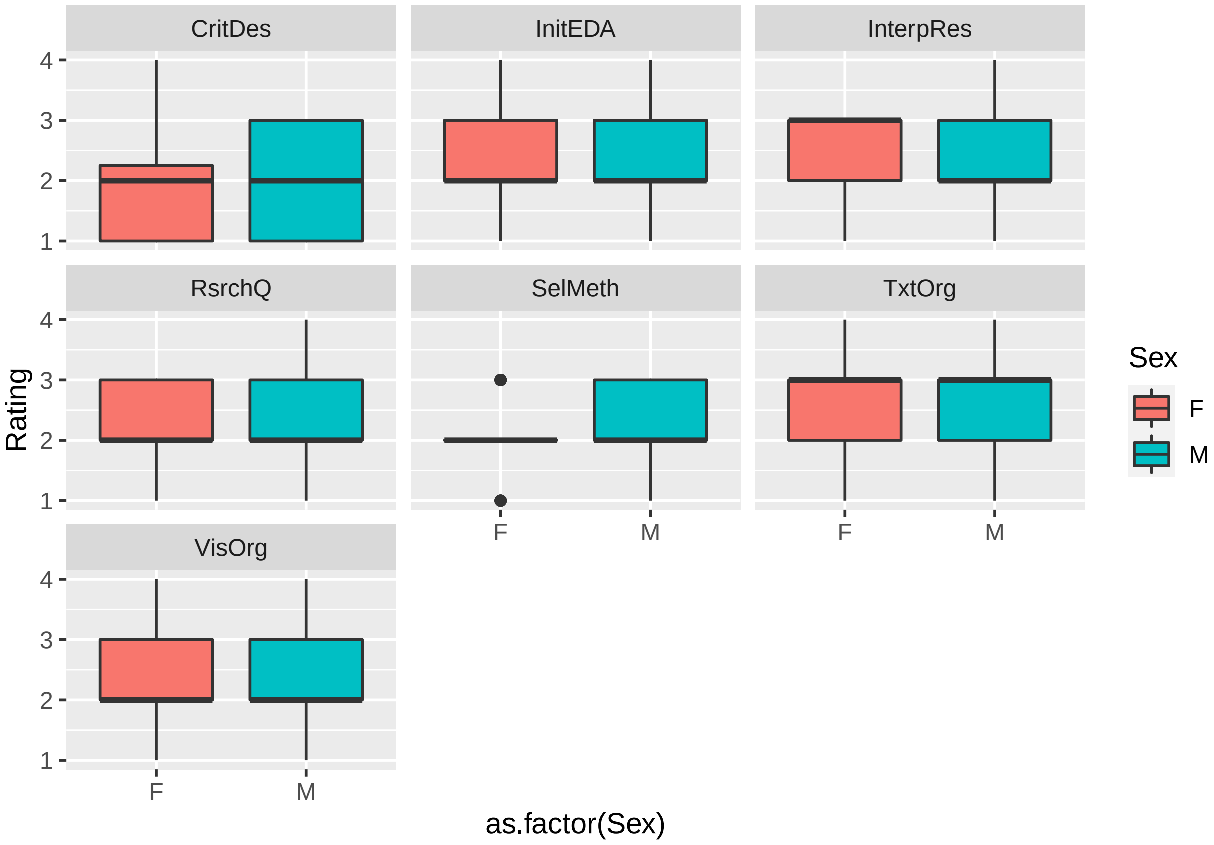
```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



We repeat the above analysis with boxplots and we can make the same conclusions about raters 2 and 3.

```
ggplot(data = tall) + geom_boxplot(aes(y = Rating, x = as.factor(Sex),
                                       fill = as.factor(Sex))) +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Sex"))
```

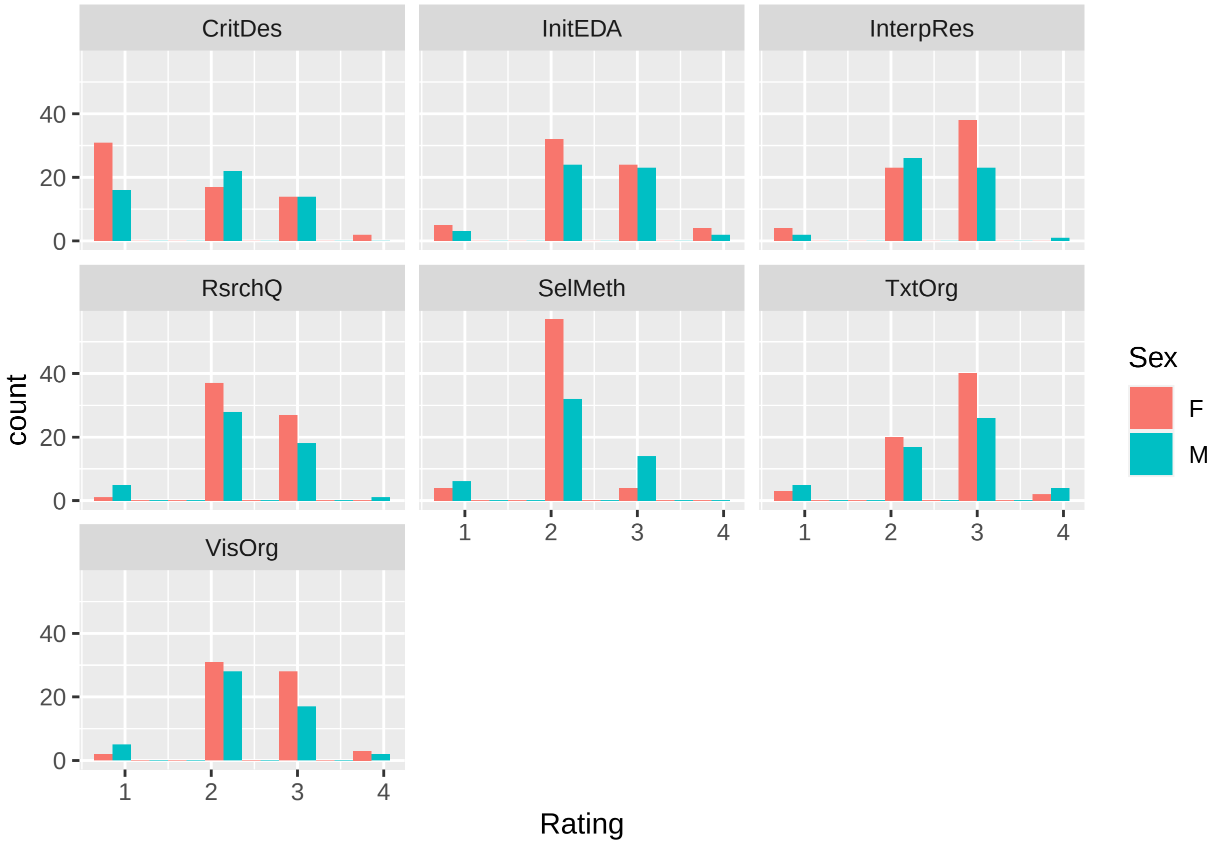
```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



Next, we group by Sex and see that the distribution of ratings in most rubrics looks similar across male and female individuals. This justifies our earlier argument that Sex does not impact rating much in any rubric, so our impputation strategy is reasonable.

```
ggplot(data = tall) + geom_histogram(aes(x = Rating, fill = as.factor(Sex)),
                                     bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Sex"))
```

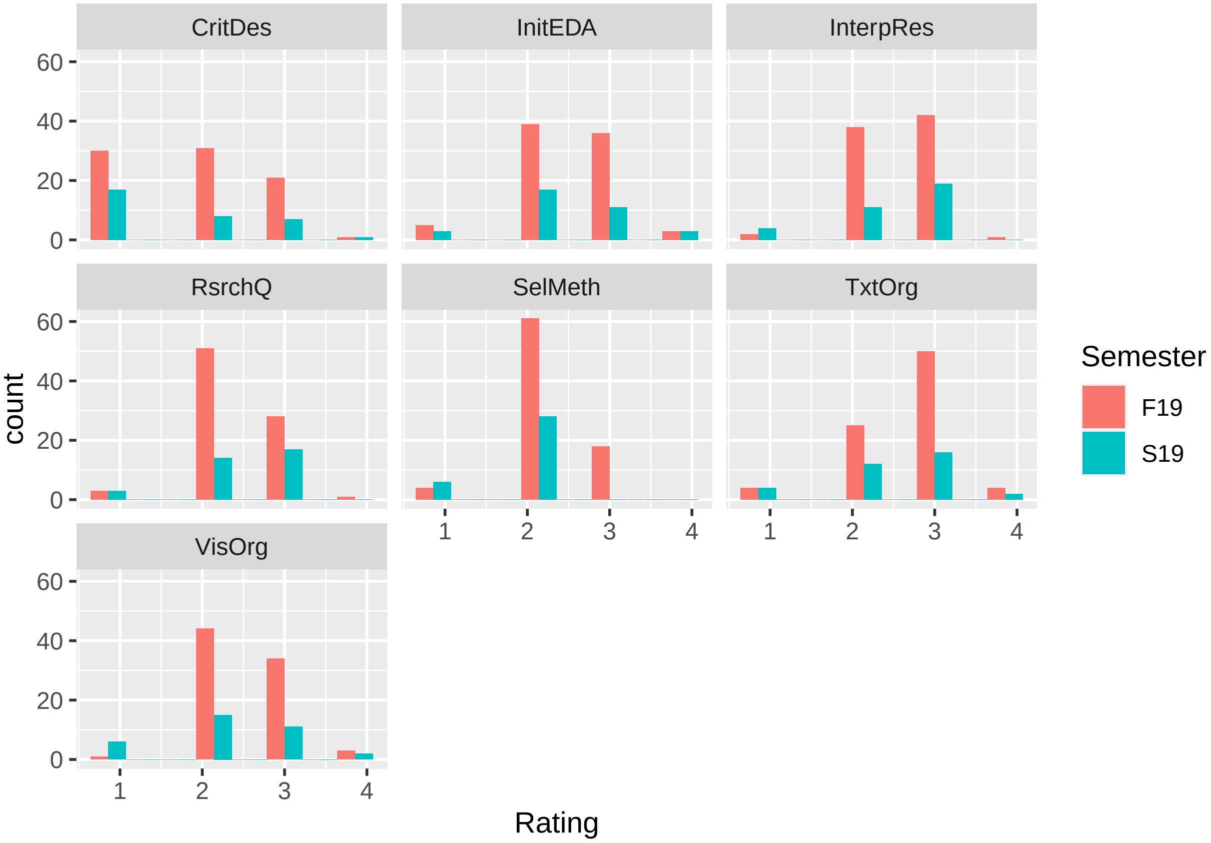
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



These histograms show only slight, noisy differences in distribution when split by Sex.

```
ggplot(data = tall) + geom_histogram(aes(x = Rating,
                                         fill = as.factor(Semester)),
                                     bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Semester"))
```

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



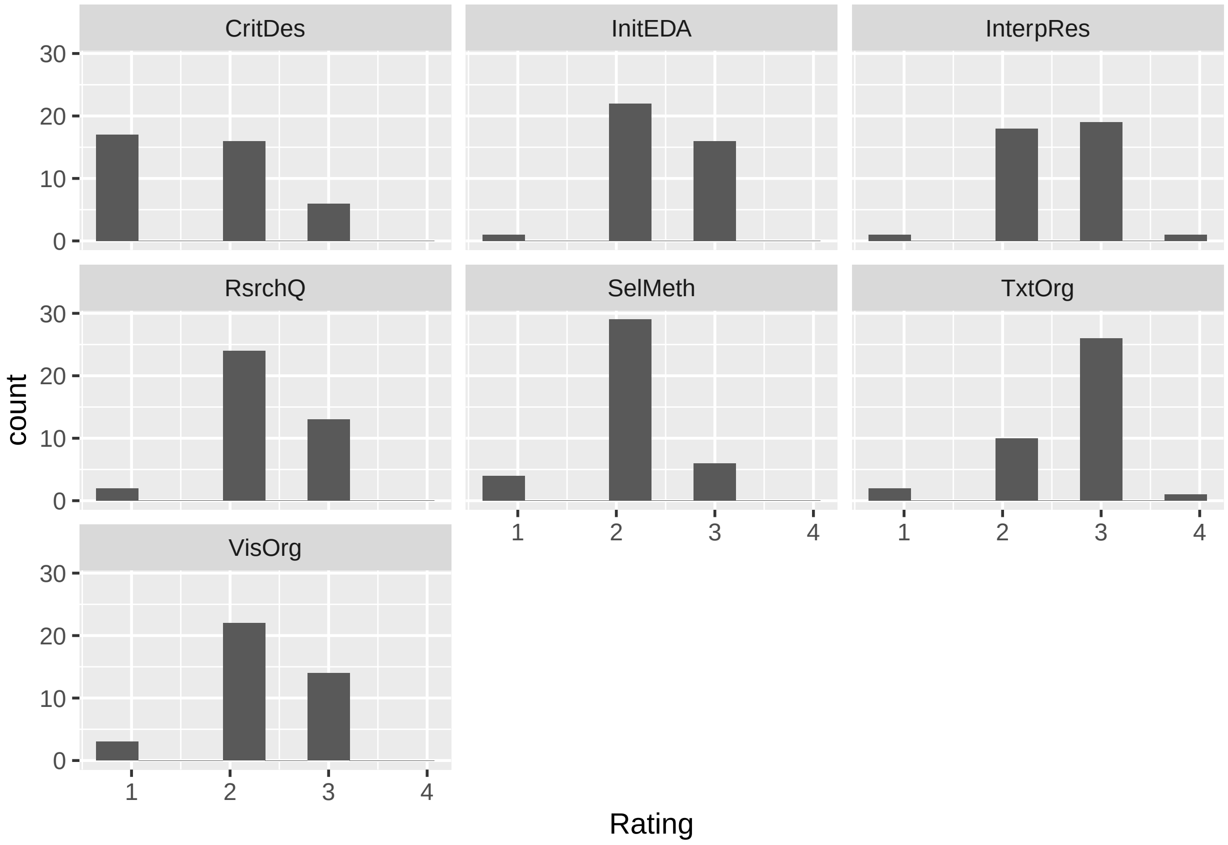
There do seem to be important differences in distribution when splitting by semester, possibly due to the fact that there were many more examples in Fall than Spring. A few examples: For sel meth, no 3s were given in spring but around 20 were given in the fall. Also, for research question, the most common rating in fall was a 2 but it was a 3 in the spring.

Overall, our EDA suggests that rater and semester can both be related to ratings, and that the distribution of rating and its relationship with these variables changes across rubrics.

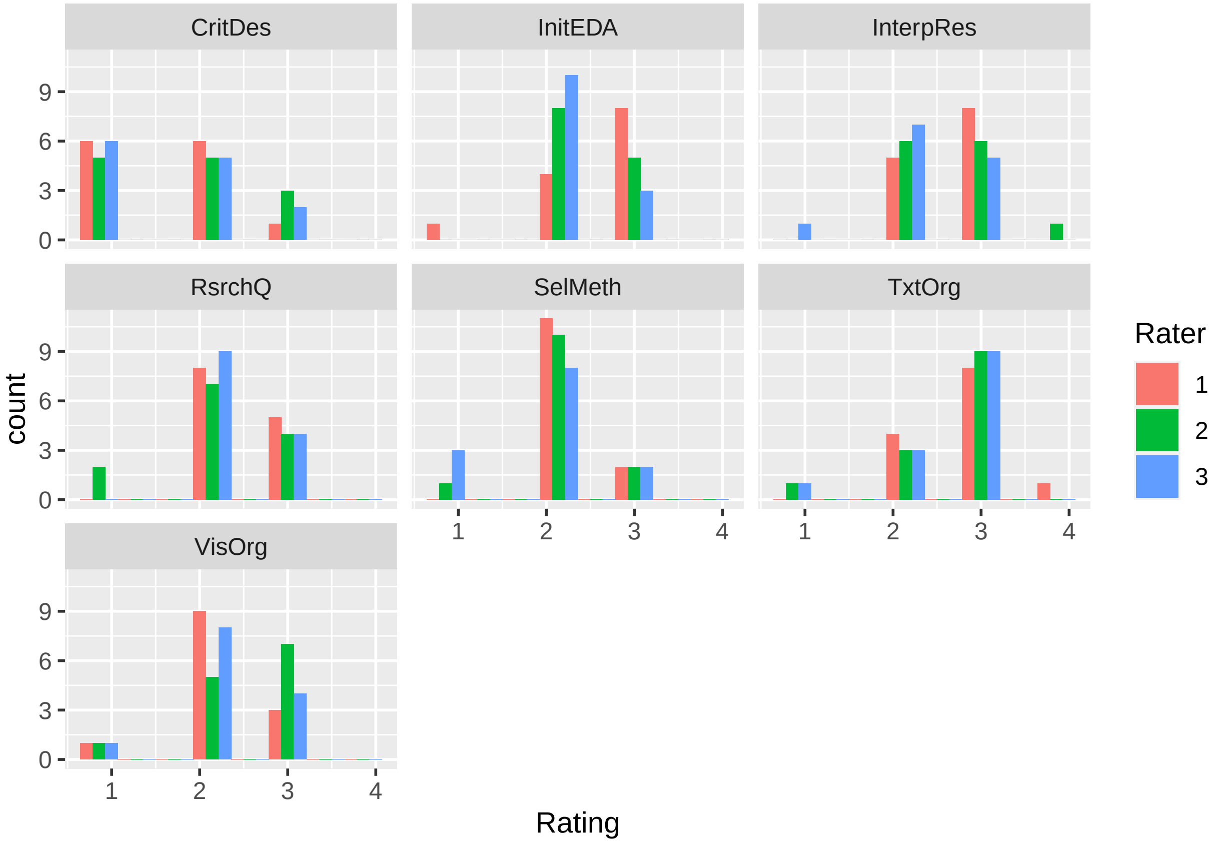
We next perform the exact same analysis using only the 13 common artifacts across the 3 raters.

```
common <- tall[grep("0",tall$Artifact),]
```

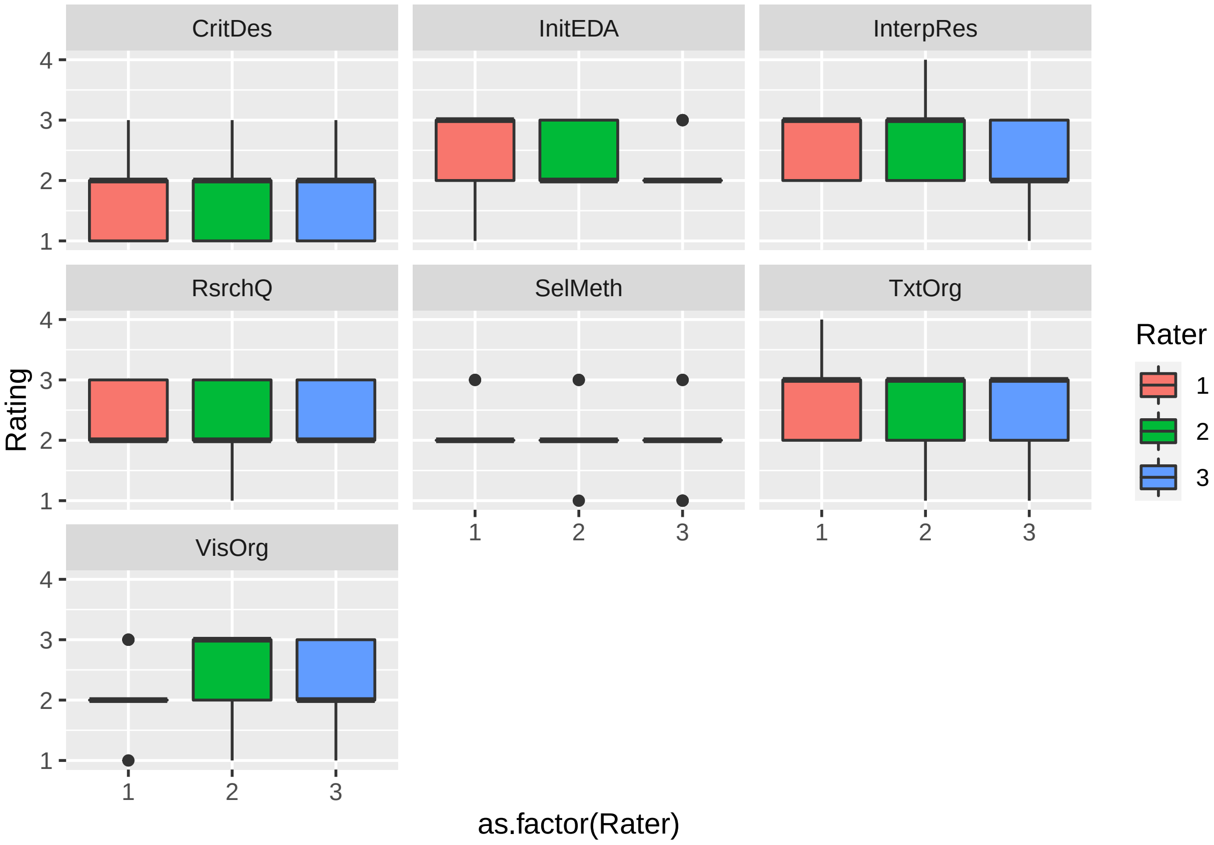
```
ggplot(data = common) + geom_histogram(aes(x = Rating),
                                       bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)
```



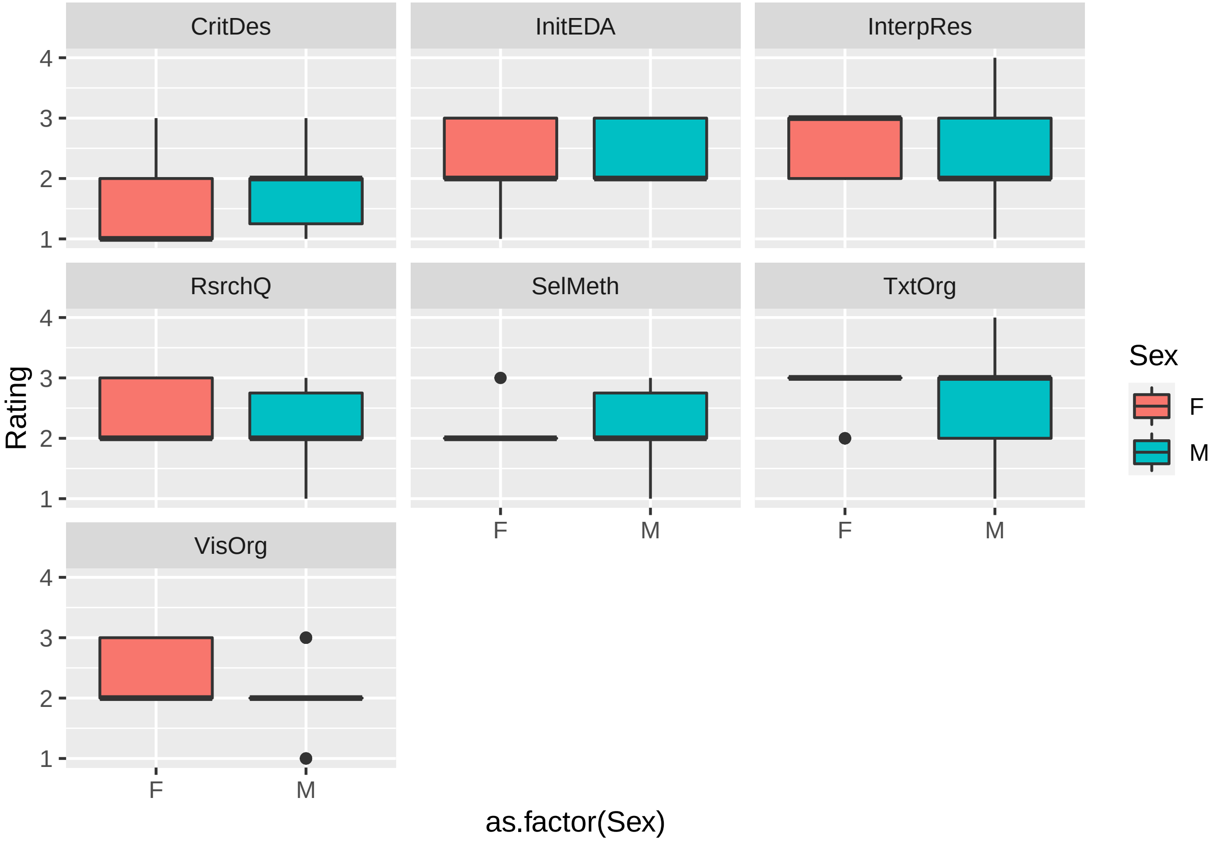
```r
ggplot(data = common) + geom_histogram(aes(x = Rating,
                                           fill = as.factor(Rater)),
                                       bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Rater"))
```



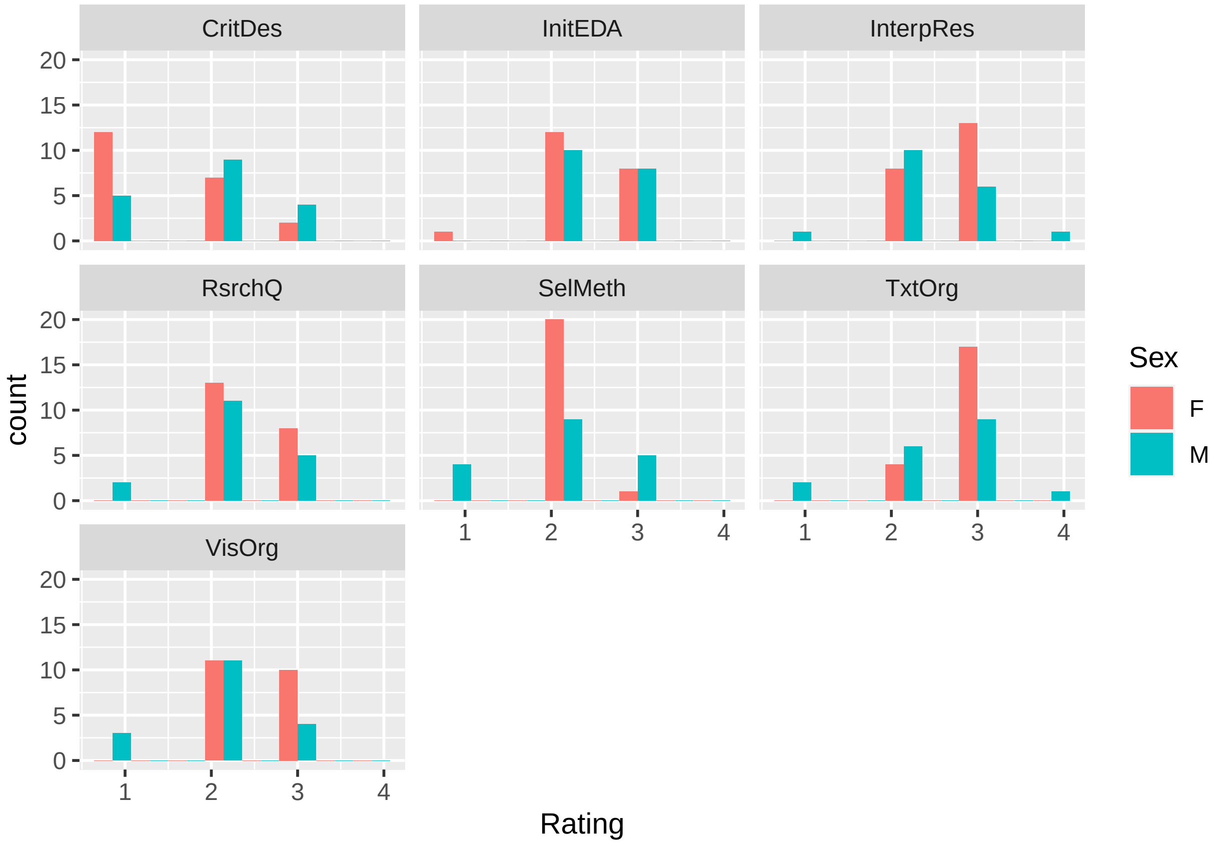
```
ggplot(data = common) + geom_boxplot(aes(y = Rating, x = as.factor(Rater),
                                        fill = as.factor(Rater))) +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Rater"))
```



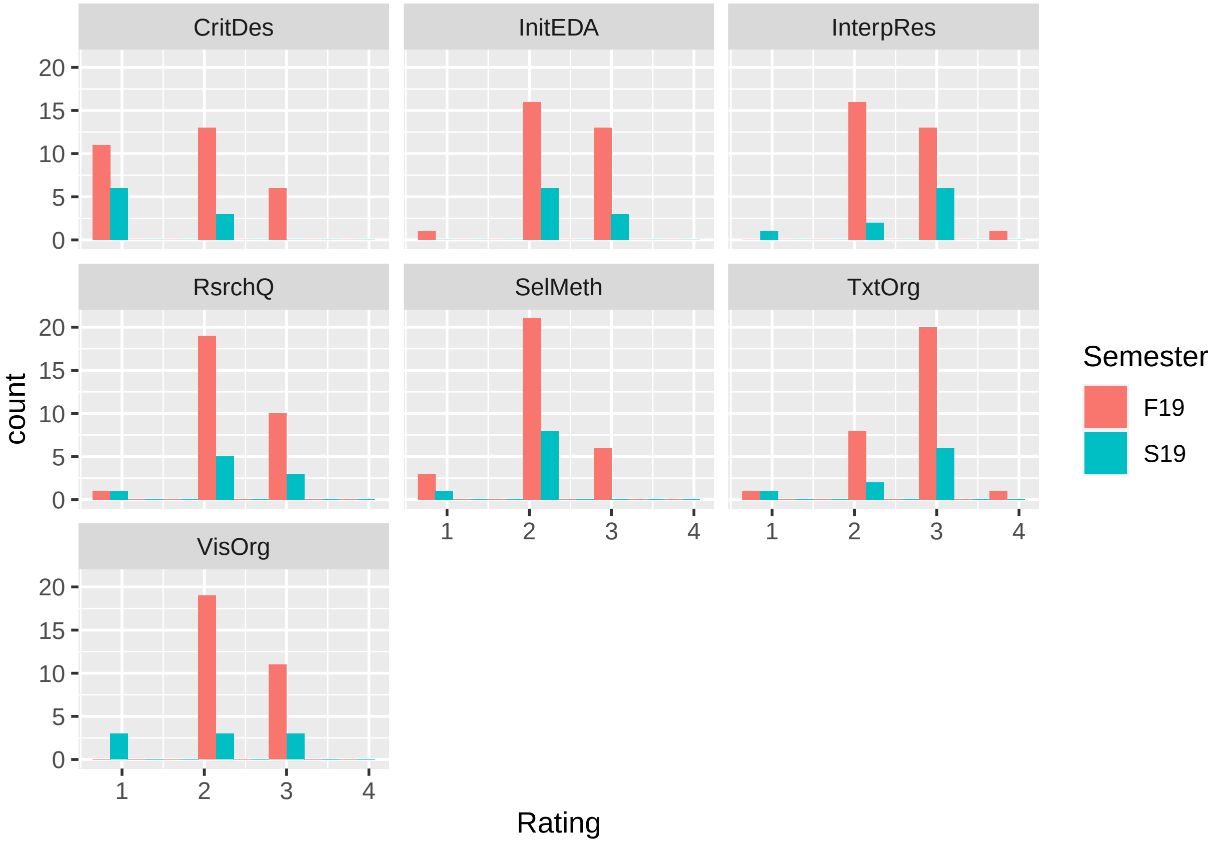
```
ggplot(data = common) + geom_boxplot(aes(y = Rating, x = as.factor(Sex),
                                         fill = as.factor(Sex))) +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Sex"))
```



```r
ggplot(data = common) + geom_histogram(aes(x = Rating, fill = as.factor(Sex)),
                                       bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Sex"))
```



```
ggplot(data = common) + geom_histogram(aes(x = Rating,
                                          fill = as.factor(Semester)),
                                      bins = 8, position = 'dodge') +
  facet_wrap(~Rubric)+ guides(fill=guide_legend(title="Semester"))
```



Overall, the conclusions we can make from these graphics are the same as those for the whole ratings dataset. This means that the 13 common samples can be taken as representative of the whole dataset, at least in approximation.

Research Question 2

Within each rubric category, do raters generally agree on their scores? Is there any pattern in rater disagreement?

Calculating ICCs

Common Data We are now interested in measuring agreement across the different raters. One way to do this is to fit a multilevel model for each rubric and regress raters against an intercept and a random artifact effect. This helps us measure correlation between raters because in each in each artifact group we have one observation for each rater, so the ICC for these models would measure the correlation between each rater across these artifacts.

```
RsrchQ.ratings <- common[common$Rubric=="RsrchQ",]
```

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

summary(lmer_RsrchQ )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 66.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3025 -0.5987 -0.3276  0.9696  1.6472
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05983  0.2446
##  Residual             0.25641  0.5064
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1057   21.59
```

```
icc_rsrch =  0.1891918
```

ICC = 0.05983/(0.05983 + 0.25641) = 0.1891918

```
CritDes.ratings <- common[common$Rubric=="CritDes",]
```

```
lmer_CritDes = lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)

summary(lmer_CritDes )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
##
## REML criterion at convergence: 75.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9647 -0.4386 -0.2978  0.5318  2.1987
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3091   0.5560
##  Residual             0.2308   0.4804
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.7179     0.1723   9.969
```

```
icc_crit = 0.5725134
```

ICC = 0.3091 / (0.3091 + 0.2308) = 0.5725134

```
InitEDA.ratings <- common[common$Rubric=="InitEDA",]
```

```
lmer_InitEDA = lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)

summary(lmer_InitEDA )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
##
## REML criterion at convergence: 56.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1670 -0.2504 -0.2504  0.4006  1.6663
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1496   0.3867
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.3846     0.1243   19.18
```

```
icc_init = 0.4930784
```

ICC = 0.1496 / (0.1496 + 0.1538) = 0.4930784.

```
SelMeth.ratings <- common[common$Rubric=="SelMeth",]
```

```
lmer_SelMeth = lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)

summary(lmer_SelMeth )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
##
## REML criterion at convergence: 50.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.11366 -0.03357 -0.03357  0.62101  2.04652
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1396   0.3736
##  Residual             0.1282   0.3581
## Number of obs: 39, groups:  Artifact, 13
```

```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.0513     0.1184   17.32
```

```
icc_sel = 0.5212845
```

ICC = 0.1396 / (0.1396 + 0.1282) = 0.5212845.

```
InterpRes.ratings <- common[common$Rubric=="InterpRes",]
```

```
lmer_InterpRes = lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)

summary(lmer_InterpRes )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
##
## REML criterion at convergence: 71.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0965 -0.8061  0.4844  0.7806  2.6635
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08405  0.2899
##  Residual             0.28205  0.5311
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    2.513      0.117   21.47
```

```
icc_interp = 0.2295821
```

ICC = 0.08405 / (0.08405 + 0.28205) = 0.2295821.

```
VisOrg.ratings <- common[common$Rubric=="VisOrg",]
```

```
lmer_VisOrg = lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)

summary(lmer_VisOrg )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
##
## REML criterion at convergence: 60.5
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5168 -0.7176 -0.1341  0.3414  1.7241
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2236   0.4729
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1454   15.69
```

```
icc_vis = 0.5924748
```

ICC = 0.2236 / (0.2236 + 0.1538) = 0.5924748.

```
TxtOrg.ratings <- common[common$Rubric=="TxtOrg",]
```

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

summary(lmer_TxtOrg )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 74.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6943 -0.7698  0.3849  0.3849  2.5019
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05556  0.2357
##  Residual             0.33333  0.5774
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6667     0.1132   23.55
```

```
icc_txt = 0.1428682
```

ICC = 0.05556 / (0.05556 + 0.33333) = 0.1428682.

```
icc_small = data.frame(Rubric = names(ratings)[7:13], ICC = c(icc_rsrch,
                                                              icc_crit,
                                                              icc_init,
                                                              icc_sel,
                                                              icc_interp,
                                                              icc_vis,
                                                              icc_txt))

library(knitr)

kable(icc_small)
```

Rubric	ICC
RsrchQ	0.1891918
CritDes	0.5725134
InitEDA	0.4930784
SelMeth	0.5212845
InterpRes	0.2295821
VisOrg	0.5924748
TxtOrg	0.1428682

Above we have the ICCs for each rubric. We can see that some rubrics have more disagreement than others. The ICC for text org and research question is lower comparatively, suggesting that rater's assesments were uncorrelated, i.e. they had a lot of disagreement. On the other hand, the ratings for crit des and sel meth are higher comparatively, suggesting that the raters tended to give similar scores for these rubrics.

Full Data We can repeat this analysis for the full data:

```
RsrchQ.ratings <- tall[tall$Rubric=="RsrchQ",]
```

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

summary(lmer_RsrchQ )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
##
## REML criterion at convergence: 211.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2748 -0.5365 -0.3780  0.9626  2.4617
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07372  0.2715
##  Residual             0.27797  0.5272
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35790    0.05774   40.84
```

```
icc_rsrch = 0.07372/(0.07372 + 0.27797)
```

ICC = 0.07372/(0.07372 + 0.27797) = 0.2096164.

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

```
lmer_CritDes = lmer(Rating ~ 1 + (1|Artifact), data=CritDes.ratings)

summary(lmer_CritDes )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: CritDes.ratings
##
## REML criterion at convergence: 277.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.01042 -0.60409  0.04407  0.72769  2.06310
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4963   0.7045
##  Residual             0.2411   0.4910
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.90720    0.08874   21.49
```

```
icc_crit = 0.4963 / (0.4963 + 0.2411)
```

ICC = 0.4963 / (0.4963 + 0.2411) = 0.6730404

```
InitEDA.ratings <- tall[tall$Rubric=="InitEDA",]
```

```
lmer_InitEDA = lmer(Rating ~ 1 + (1|Artifact), data=InitEDA.ratings)

summary(lmer_InitEDA )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
##
## REML criterion at convergence: 240.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8923 -0.3451 -0.1454  0.4250  1.6015
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3628   0.6023
##  Residual             0.1655   0.4068
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44815    0.07479   32.73
```

```
icc_init = 0.3628/(0.3628 +0.1655)
```

ICC = 0.3628 / (0.3628 + 0.1655) = 0.686731.

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

```
lmer_SelMeth = lmer(Rating ~ 1 + (1|Artifact), data=SelMeth.ratings)

summary(lmer_SelMeth )
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: SelMeth.ratings
##
## REML criterion at convergence: 157.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2057 -0.1075 -0.1075 -0.0553  2.0951
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1108   0.3329
##  Residual             0.1240   0.3521
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.07168    0.04893   42.34
```

```
icc_sel = .1108/(.1108 +0.1240)
```

ICC = 0.1108 / (0.1108 + 0.1240) = 0.471891.

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

```
lmer_InterpRes = lmer(Rating ~ 1 + (1|Artifact), data=InterpRes.ratings)

summary(lmer_InterpRes )
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InterpRes.ratings
##
## REML criterion at convergence: 217.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1448 -0.6998  0.5175  0.7452  2.6532
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08219  0.2867
##  Residual             0.29136  0.5398
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.48427    0.05962   41.67
```

```
icc_interp = 0.08219/(0.08219+0.29136)
```

ICC = 0.08219 / (0.08219 + 0.29136) = 0.2200241.

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

```
lmer_VisOrg = lmer(Rating ~ 1 + (1|Artifact), data=VisOrg.ratings)

summary(lmer_VisOrg )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: VisOrg.ratings
##
## REML criterion at convergence: 226.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5918 -0.3789 -0.1632  0.4726  1.6322
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3092   0.5561
##  Residual             0.1588   0.3985
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44497    0.07063   34.62
```

```
icc_vis = .3092/(.3092 + 0.1588)
```

ICC = 0.3092 / (0.3092 + 0.1588) = 0.6606838.

```
TxtOrg.ratings <- tall[tall$Rubric=="TxtOrg",]
```

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

summary(lmer_TxtOrg )
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 249
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3638 -0.7641  0.3836  0.5278  2.4094
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09145  0.3024
##  Residual             0.39503  0.6285
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.59144    0.06764   38.31
```

```
icc_txt = 0.09145/(0.09145 + 0.39503)
```

```
icc_big = data.frame(Rubric = names(ratings)[7:13], ICC_all = c(icc_rsrch,
                                                                icc_crit,
                                                                icc_init,
                                                                icc_sel,
                                                                icc_interp,
                                                                icc_vis,
                                                                icc_txt))

library(knitr)

kable(cbind(icc_big, icc_small))
```

Rubric	ICC_all	Rubric	ICC
RsrchQ	0.2096164	RsrchQ	0.1891918
CritDes	0.6730404	CritDes	0.5725134
InitEDA	0.6867310	InitEDA	0.4930784
SelMeth	0.4718910	SelMeth	0.5212845
InterpRes	0.2200241	InterpRes	0.2295821
VisOrg	0.6606838	VisOrg	0.5924748
TxtOrg	0.1879831	TxtOrg	0.1428682

We see in general that the ICCs are similar across the common and full data, indicating again that the 13 common artifacts are a reasonable representation of the full dataset.

Calculating Exact Agreement Because we have evidence that we can look only at the 13 common artifacts to measure agreement, we now calculate the exact agreement between raters for each rubric as follows below. The table presents results for each possible pair of raters / rubric group and the percent exact agreement between those two raters for all artifact scores in that rubric category.

```
ratings_small = ratings[ratings$Repeated == 1 ,]
```

```
pairs = combn(1:3, 2)
rubrics = names(ratings)[7:13]
d = data.frame(first = NULL, second = NULL, Rubric = Null, pct.exact = NULL)
for(i in 1:ncol(pairs)){
  pair = pairs[,i]
  first = pair[1]
  second = pair[2]
  for(rubric in rubrics){
    t = table(ratings_small[ratings_small$Rater == first, rubric],
              ratings_small[ratings_small$Rater == second, rubric])
    minrow = as.numeric(min(rownames(t)))
    mincol = as.numeric(min(colnames(t)))
    if(minrow > mincol){

      t2 = t[1:nrow(t), max(c(minrow, mincol)):ncol(t)]
      #print(max(c(minrow, mincol)))
      #print(c(first, second, rubric, sum(diag(t2))/sum(t2)))
      d = rbind(d , c(first, second, rubric, sum(diag(t2))/sum(t)))
      #print(t2)
    }
    else if(minrow < mincol){
      t2 = t[max(c(minrow, mincol)):nrow(t),1:ncol(t)]
      #print(max(c(minrow, mincol)))
      #print(c(first, second, rubric, sum(diag(t2))/sum(t2)))
      d = rbind(d , c(first, second, rubric, sum(diag(t2))/sum(t)))
      #print(t2)
    }
    else{
      #print(c(first, second, rubric, sum(diag(t))/sum(t)))
      d = rbind(d , c(first, second, rubric, sum(diag(t))/sum(t)))
      #print(t)
    }

  }
}

d[,4] = as.numeric(d[,4])
names(d) = c('First', 'Second', 'Rubric', 'Percent Exact Agreement')
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x tidyr::expand()    masks Matrix::expand()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x tidyr::pack()      masks Matrix::pack()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::select()    masks MASS::select()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
## x tidyr::unpack()    masks Matrix::unpack()
```

```
d$`Percent Exact Agreement` = round(d$`Percent Exact Agreement`, 2)
kable(d %>% pivot_wider(names_from = 'Rubric',
                        values_from = `Percent Exact Agreement`))
```

First	Second	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
1	2	0.38	0.54	0.69	0.92	0.62	0.54	0.69
1	3	0.77	0.62	0.54	0.62	0.54	0.77	0.62
2	3	0.54	0.69	0.85	0.69	0.62	0.77	0.54

In general, it seems like raters 1 and 3 tend to agree less for a lot of rubrics, while any pair with rater 2 seems to have a slightly higher agreement. This makes sense, as rater 3 is harsh, rater 1 is lenient, and rater 2 seems to be in between them. There are a couple strange observations, i.e. the 90% agreement between raters 1 and 2 for sel meth but only 20% agreement for research question. This suggests that there is a lot of variability in rater agreement across rubric.

Research Question 3

How do other factors, like student sex, semester the course was taken, etc. affect the ratings?

A natural follow-up to Question 2 is to ask what other variables relate to the ratings and how exactly do these variables interact? Our process for answering this question is to proceed with the multilevel models from 2, add fixed effects and a few random effects, perform variable selection, recompute ICC, and determine what variables improved the fit. Note, not all fixed / random effects we tried are shown here because they were not all useful.

Model Selection

Common Data We first try this process on the small data with only 13 artifacts, and then for the full data. We start with automatic variable selection from the HW10 solutions, and then validate the results manually with ANOVA and exploration of other possible random effects.

```r
library(lme4)
library(LMERConvenienceFunctions)
tall.13 <- tall[grep("O",tall$Artifact),]
ICC.vec.small = c()
Rubric.names <- sort(unique(tall$Rubric))
model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- Rubric.names
## There will be a lot of output from fitLMER.fnc() here... Sorry!
for (i in Rubric.names) {
## fit each base model
rubric.data <- tall.13[tall.13$Rubric==i,]
tmp <- lme4::lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
Semester + Sex + (1|Artifact),
data=rubric.data,REML=FALSE)
## do backwards elimination
tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
## check to see if the raters are significantly different from one another
tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
## choose the best model
if (pval<=0.05) {
tmp_final <- tmp.back_elim
} else {
tmp_final <- tmp.single_intercept
}
sig2 <- summary(tmp_final)$sigma^2
tau2 <- attr(summary(tmp_final)$varcor[[1]],"stddev")^2
ICC <- tau2 / (tau2 + sig2)
ICC.vec.small <- c(ICC.vec.small,ICC)

## and add to list...
model.formula.13[[i]] <- formula(tmp_final)
}
```

```
## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE
```

```
## ======================================================
## ===             backfitting fixed effects          ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.2229 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1826 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8137 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.6429 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.8294 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.2947 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.7355 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.279 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.9383 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.4287 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.5358 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1319 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.1922 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.1078 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

```
model.formula.13
```

```
## $CritDes
## as.numeric(Rating) ~ (1 | Artifact)
##
## $InitEDA
## as.numeric(Rating) ~ (1 | Artifact)
```

```
##
## $InterpRes
## as.numeric(Rating) ~ (1 | Artifact)
##
## $RsrchQ
## as.numeric(Rating) ~ (1 | Artifact)
##
## $SelMeth
## as.numeric(Rating) ~ (1 | Artifact)
##
## $TxtOrg
## as.numeric(Rating) ~ (1 | Artifact)
##
## $VisOrg
## as.numeric(Rating) ~ (1 | Artifact)
```

For the small data, we get that only the random intercept term is significant for every rubric item. That is, at least for this data, the other variables do not seem to matter with regards to modeling rating.

```r
library(lme4)
library(LMERConvenienceFunctions)
tall.nonmissing <- tall
model.formula.alldata <- as.list(rep(NA,7))
model.coef.alldata <- as.list(c())
model.ranef.alldata <- as.list(c())
Rubric.names <- sort(unique(tall$Rubric))
names(model.formula.alldata) <- Rubric.names
ICC.Vec.all = c()
## There will be a lot of output from fitLMER.fnc() here... Sorry!
for (i in Rubric.names) {
  ## fit each base model
  rubric.data <- tall.nonmissing[tall.nonmissing$Rubric==i,]
  tmp <- lme4::lmer(as.numeric(Rating) ~ -1 + as.factor(Rater) +
  Semester + Sex + (1|Artifact),
  data=rubric.data,REML=FALSE)
  ## do backwards elimination
  tmp.back_elim <- fitLMER.fnc(tmp,set.REML.FALSE = TRUE,log.file.name = FALSE)
  ## check to see if the raters are significantly different from one another
  tmp.single_intercept <- update(tmp.back_elim, . ~ . + 1 - as.factor(Rater))
  pval <- anova(tmp.single_intercept,tmp.back_elim)$"Pr(>Chisq)"[2]
  ## choose the best model
  if (pval<=0.05) {
    tmp_final <- tmp.back_elim
  } else {
    tmp_final <- tmp.single_intercept
  }
  sig2 <- summary(tmp_final)$sigma^2
  tau2 <- attr(summary(tmp_final)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.Vec.all <- c(ICC.Vec.all,ICC)
  ## and add to list...
  model.formula.alldata[[i]] <- formula(tmp_final)
  model.coef.alldata[[i]] <- summary(tmp_final)$coef
  model.ranef.alldata[[i]] <- summary(tmp_final)$varcor
}
```

Full Data

```
## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE
```

```
## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.7022 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.6521 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune
```

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

```
## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE
```

```
## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.8529 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.83 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.501 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.473 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE
## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.5212 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.4453 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.3095 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune
## refitting model(s) with ML (instead of REML)

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Sex" = 0.4508 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Semester" = 0.1874 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune

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

## Warning in fitLMER.fnc(tmp, set.REML.FALSE = TRUE, log.file.name = FALSE): Argument "ran.effects" is
## TRUE

## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## processing model terms of interaction level 1
##   iteration 1
##     p-value for term "Semester" = 0.1902 >= 0.05
##     not part of higher-order interaction
##     removing term
##   iteration 2
##     p-value for term "Sex" = 0.3046 >= 0.05
##     not part of higher-order interaction
##     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
## pruning random effects structure ...
##   nothing to prune
```

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

```
model.formula.alldata
```

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

However, when performing the same experiment for the full data, we find that rater and semester appear as important fixed effects for certain rubrics. We used no global intercept in the automated fitting to make the process easier for the fitting algorithm, so to validate our results when including a global intercept we perform manual inspection and variable selection below.

```
RsrchQ.ratings[5, 'Sex'] = 'F'
mlm1_RsrchQ = lmer(Rating ~ 1 + (Semester + Sex  + Repeated) * Rater +
                     (1|Artifact), data=RsrchQ.ratings, REML=F)

mlm2_RsrchQ = lmer(Rating ~ 1  + Rater + (1|Artifact), data=RsrchQ.ratings,
                   REML=F)
BIC(lmer_RsrchQ)
```

```
## [1] 225.3524
```

```
BIC(mlm2_RsrchQ)
```

```
## [1] 229.2031
```

```
BIC(mlm1_RsrchQ)
```

```
## [1] 265.8773
```

```
summary(mlm2_RsrchQ)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: Rating ~ 1 + Rater + (1 | Artifact)
##    Data: RsrchQ.ratings
##
##      AIC      BIC   logLik deviance df.resid
##    215.4    229.2   -102.7    205.4      112
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2947 -0.5454 -0.4175  0.8706  2.3845
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06678  0.2584
##  Residual             0.27595  0.5253
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44422    0.09302  26.276
## Rater2      -0.08841    0.12753  -0.693
## Rater3      -0.17183    0.12753  -1.347
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.685
## Rater3 -0.685  0.500
```

```
summary(lmer_RsrchQ)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: RsrchQ.ratings
## 
## REML criterion at convergence: 211.1
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.2748 -0.5365 -0.3780  0.9626  2.4617 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07372  0.2715  
##  Residual             0.27797  0.5272  
## Number of obs: 117, groups:  Artifact, 91
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35790    0.05774   40.84
```

```
anova(mlm2_RsrchQ, lmer_RsrchQ)
```

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

```
## Data: RsrchQ.ratings
## Models:
## lmer_RsrchQ: Rating ~ 1 + (1 | Artifact)
## mlm2_RsrchQ: Rating ~ 1 + Rater + (1 | Artifact)
##             npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## lmer_RsrchQ    3 213.19 221.48 -103.6   207.19                     
## mlm2_RsrchQ    5 215.39 229.20 -102.7   205.39 1.8013  2     0.4063
```

```
icc_rsrch2 = 0.07372  / (0.07372  + 0.27797)
icc_rsrch3 = 0.05539 / (0.05539 + 0.27429)
```

For research question, we performed a lot of manual variable selection with BIC as a criterion for model improvement. However, no model we tried with either fixed or random effects led to any improvement. In fact, the only model that did not worsen the BIC was a simple fixed effect for rater in addition to the artifact random effect. However, as we can see from the anova output above, this fixed effect (along with the others, not shown) are not significant. We recomputed ICC on both a full model with all fixed effects and interactions with rater, and the reduced one with just a fixed effect for rater. We also experimented with other interactions and deeper levels of interactions, but nothing ended up being meaningful.

```
CritDes.ratings[5, 'Sex'] = 'F'
mlm1_CritDes = lmer(Rating ~ 1 + (Sex + Semester  + Repeated) * Rater +
                      (1|Artifact), data=CritDes.ratings, REML=F)

mlm2_CritDes = lmer(Rating ~ 1 +Rater + (1|Artifact), data=CritDes.ratings,
                    REML=F)
BIC(lmer_CritDes)
```

```
## [1] 292.1299
```

```
BIC(mlm2_CritDes)
```

```
## [1] 290.6254
```

```
anova(lmer_CritDes, mlm2_CritDes)
```

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

## Data: CritDes.ratings
## Models:
## lmer_CritDes: Rating ~ 1 + (1 | Artifact)
## mlm2_CritDes: Rating ~ 1 + Rater + (1 | Artifact)
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_CritDes    3 280.86 289.12 -137.43   274.86
## mlm2_CritDes    5 276.86 290.62 -133.43   266.86 7.9996  2    0.01832 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
BIC(mlm1_CritDes)
```

```
## [1] 326.5742
```

```
summary(mlm2_CritDes)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: Rating ~ 1 + Rater + (1 | Artifact)
##    Data: CritDes.ratings
##
##      AIC      BIC   logLik deviance df.resid
##    276.9    290.6   -133.4    266.9      111
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.59134 -0.50054 -0.08452  0.63588  1.65959
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4381   0.6619
##  Residual             0.2355   0.4852
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   1.6948     0.1192  14.213
## Rater2        0.4225     0.1466   2.882
## Rater3        0.2194     0.1460   1.502
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.606
## Rater3 -0.612  0.498
```

```
kable(round(summary(mlm2_CritDes)$coef, 2))
```

	Estimate	Std. Error	t value
(Intercept)	1.69	0.12	14.21
Rater2	0.42	0.15	2.88
Rater3	0.22	0.15	1.50

```
kable((summary(mlm2_CritDes)$varcor))
```

grp	var1	var2	vcov	sdcor
Artifact	(Intercept)	NA	0.4381022	0.6618929
Residual	NA	NA	0.2354533	0.4852353

```
icc_crit2 = 0.4381 / (0.4381 +0.2355 )
icc_crit3 = 0.4465 / (0.4465 + 0.2419)
```

We repeated the procedure for crit des, and got the same results where no meaningful fixed or random effects were found.

```
InitEDA.ratings[5, 'Sex'] = 'F'
mlm1_InitEDA = lmer(Rating ~ 1 + (Sex * Semester  * Repeated) * Rater +
                      (1|Artifact), data=InitEDA.ratings, REML=F)

mlm2_InitEDA = lmer(Rating ~ 1 + as.factor(Rater) + (1|Artifact),
                    data=InitEDA.ratings, REML=F)
BIC(lmer_InitEDA)
```

```
## [1] 255.0628
```

```
BIC(mlm2_InitEDA)
```

```
## [1] 258.0923
```

```
BIC(mlm1_InitEDA)
```

```
## [1] 338.0298
```

```
anova(lmer_InitEDA, mlm2_InitEDA)
```

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

## Data: InitEDA.ratings
## Models:
## lmer_InitEDA: Rating ~ 1 + (1 | Artifact)
## mlm2_InitEDA: Rating ~ 1 + as.factor(Rater) + (1 | Artifact)
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_InitEDA    3 243.42 251.71 -118.71   237.42
## mlm2_InitEDA    5 244.28 258.09 -117.14   234.28 3.1408  2      0.208
```

```
summary(mlm2_InitEDA)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + as.factor(Rater) + (1 | Artifact)
##    Data: InitEDA.ratings
## 
##      AIC      BIC   logLik deviance df.resid 
##    244.3    258.1   -117.1    234.3      112 
## 
## Scaled residuals: 
##      Min       1Q   Median       3Q      Max 
## -2.12233 -0.37243 -0.01405  0.36506  1.55569 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3730   0.6107  
##  Residual             0.1471   0.3836  
## Number of obs: 117, groups:  Artifact, 91
## 
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        2.50499    0.10174  24.622
## as.factor(Rater)2  0.01293    0.12024   0.107
## as.factor(Rater)3 -0.18261    0.12024  -1.519
## 
## Correlation of Fixed Effects:
##             (Intr) a.(R)2
## as.fctr(R)2 -0.591       
## as.fctr(R)3 -0.591  0.500
```

```
summary(lmer_InitEDA)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: InitEDA.ratings
## 
## REML criterion at convergence: 240.8
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -1.8923 -0.3451 -0.1454  0.4250  1.6015 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3628   0.6023  
##  Residual             0.1655   0.4068  
## Number of obs: 117, groups:  Artifact, 91
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44815    0.07479   32.73
```

```
icc_init2 = 0.3628 / (0.3628 + 0.1655)
icc_init3 = 0.3793/ (0.3793 + .1229)
```

We see the same pattern for initial EDA.

```
SelMeth.ratings[5, 'Sex'] = 'F'
library(lme4)
mlm1_SelMeth = lme4::lmer(Rating ~ 1 + (Sex + Semester  + Repeated) * Rater +
                            (1|Artifact), data=SelMeth.ratings, REML=F)

mlm2_SelMeth = lme4::lmer(Rating ~ 1 + as.factor(Rater)  + Semester +
                            (1|Artifact), data=SelMeth.ratings, REML=F)

mlm3_SelMeth = lme4::lmer(Rating ~ 1  + Semester +
                            (1|Artifact) , data=SelMeth.ratings, REML=F)
BIC(lmer_SelMeth)
```

```
## [1] 172.024
```

```
BIC(mlm2_SelMeth)
```

```
## [1] 163.4941
```

```
BIC(mlm1_SelMeth)
```

```
## [1] 191.6827
```

```
BIC(mlm3_SelMeth)
```

```
## [1] 159.6926
```

```
anova(lmer_SelMeth, mlm2_SelMeth)
```

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

```
## Data: SelMeth.ratings
## Models:
## lmer_SelMeth: Rating ~ 1 + (1 | Artifact)
## mlm2_SelMeth: Rating ~ 1 + as.factor(Rater) + Semester + (1 | Artifact)
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_SelMeth    3 159.53 167.82 -76.768   153.53
## mlm2_SelMeth    6 146.92 163.49 -67.461   134.92 18.614  3  0.0003285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kable(round(summary(mlm3_SelMeth)$coef, 2))
```

	Estimate	Std. Error	t value
(Intercept)	2.18	0.05	39.97
SemesterS19	-0.37	0.10	-3.70

```
kable(summary(mlm3_SelMeth)$varcor)
```

grp	var1	var2	vcov	sdcor
Artifact	(Intercept)	NA	0.0887949	0.2979847
Residual	NA	NA	0.1172663	0.3424417

```
summary(mlm3_SelMeth)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Semester + (1 | Artifact)
##    Data: SelMeth.ratings
##
##      AIC      BIC   logLik deviance df.resid
##    148.6    159.7    -70.3    140.6      113
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4072 -0.3032 -0.1629  0.3084  2.0815
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08879  0.2980
##  Residual             0.11727  0.3424
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.18247    0.05460  39.970
## SemesterS19 -0.36807    0.09944  -3.701
##
## Correlation of Fixed Effects:
##             (Intr)
## SemesterS19 -0.549
```

```
icc_sel2 = 0.08879 / (0.08879 + 0.11727)
icc_sel3 = 0.11166 / (0.11166 + 0.08125)
```

However, for sel meth, we actually see an improvement with fixed effects. Our best model in terms of BIC included fixed effects for semester and rater with no interactions or new random effects. Again, the random effects we experimented with did not turn out to be meaningful, especially when treating rater as a factor variable.

```
InterpRes.ratings[5, 'Sex'] = 'F'
mlm1_InterpRes = lmer(Rating ~ 1 + (Sex + Rater  + Semester) * Rater +
                        (1|Artifact), data=InterpRes.ratings, REML=F)
```

```
mlm2_InterpRes = lmer(Rating ~ 1 + (  Rater  )  + (1|Artifact),
                      data=InterpRes.ratings, REML=F)

mlm3_InterpRes = lmer(Rating ~ 1 + (  Rater  )  + (1|Artifact),
                      data=InterpRes.ratings, REML=F)


BIC(lmer_InterpRes)
```

```
## [1] 232.1896
```

```
BIC(mlm2_InterpRes)
```

```
## [1] 217.4736
```

```
BIC(mlm3_InterpRes)
```

```
## [1] 217.4736
```

```
anova(mlm2_InterpRes,lmer_InterpRes)
```

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

```
## Data: InterpRes.ratings
## Models:
## lmer_InterpRes: Rating ~ 1 + (1 | Artifact)
## mlm2_InterpRes: Rating ~ 1 + (Rater) + (1 | Artifact)
##                npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## lmer_InterpRes    3 220.09 228.38 -107.048   214.09
## mlm2_InterpRes    5 203.66 217.47  -96.831   193.66 20.433  2  3.657e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
BIC(mlm1_InterpRes)
```

```
## [1] 241.1885
```

```
summary(mlm3_InterpRes)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + (Rater) + (1 | Artifact)
##    Data: InterpRes.ratings
##
##      AIC      BIC   logLik deviance df.resid
##    203.7    217.5    -96.8    193.7      112
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5375 -0.7549  0.3770  0.6604  2.6856
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06404  0.2531
##  Residual             0.24643  0.4964
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.70496    0.08845  30.581
## Rater2      -0.11798    0.12105  -0.975
## Rater3      -0.54366    0.12105  -4.491
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.684
## Rater3 -0.684  0.500
```

```
kable(round(summary(mlm3_InterpRes)$coef, 2))
```

	Estimate	Std. Error	t value
(Intercept)	2.70	0.09	30.58
Rater2	-0.12	0.12	-0.97
Rater3	-0.54	0.12	-4.49

```
kable(summary(mlm3_InterpRes)$varcor)
```

grp	var1	var2	vcov	sdcor
Artifact	(Intercept)	NA	0.0640407	0.2530626
Residual	NA	NA	0.2464255	0.4964126

```
icc_interp2 = 0.06404/ (0.06404 + 0.24643)
icc_interp3 = 0.0575 / (0.0575 + 0.2506)
```

For interp res we find a fixed effect for just rater to be meaningful in improving the model fit.

```
VisOrg.ratings[5, 'Sex'] = 'F'
mlm1_VisOrg = lmer(Rating ~ 1 + (Sex + Semester  + Repeated) * Rater +
                     (1|Artifact), data=VisOrg.ratings, REML=F)
library(lme4)
mlm2_VisOrg = lme4::lmer(Rating ~ 1 + (Rater)  +(1|Artifact),
                         data=VisOrg.ratings, REML=T)



BIC(lmer_VisOrg)
```

```
## [1] 240.678
```

```
BIC(mlm2_VisOrg)
```

```
## [1] 245.5305
```

```
BIC(mlm1_VisOrg)
```

```
## [1] 267.5955
```

```
summary(mlm2_VisOrg)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (Rater) + (1 | Artifact)
##    Data: VisOrg.ratings
##
## REML criterion at convergence: 221.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5008 -0.3334 -0.2599  0.4108  1.8726
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2937   0.5420
##  Residual             0.1454   0.3813
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.38148    0.09652  24.673
## Rater2       0.27121    0.11645   2.329
## Rater3      -0.08213    0.11645  -0.705
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.611
## Rater3 -0.611  0.504
```

```
kable(round(summary(mlm2_VisOrg)$coef, 2))
```

	Estimate	Std. Error	t value
(Intercept)	2.38	0.10	24.67
Rater2	0.27	0.12	2.33
Rater3	-0.08	0.12	-0.71

```
kable(summary(mlm2_VisOrg)$varcor)
```

grp	var1	var2	vcov	sdcor
Artifact	(Intercept)	NA	0.2937416	0.5419793
Residual	NA	NA	0.1453580	0.3812585

```r
anova(mlm2_VisOrg, lmer_VisOrg)
```

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

```
## Data: VisOrg.ratings
## Models:
## lmer_VisOrg: Rating ~ 1 + (1 | Artifact)
## mlm2_VisOrg: Rating ~ 1 + (Rater) + (1 | Artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_VisOrg    3 228.95 237.21 -111.47   222.95
## mlm2_VisOrg    5 222.97 236.74 -106.48   212.97 9.9784  2   0.006811 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
icc_vis2 = 0.2894 / (0.2894 + 0.1462)
icc_vis3 = 0.2411 / (0.2411 + 0.1685)
```

For vis org we once again find nothing relevant for prediction across many fixed and random effects we tried.

```r
TxtOrg.ratings[5, 'Sex'] = 'F'
mlm1_TxtOrg = lmer(Rating ~ 1 + (Sex + Semester  + Repeated) * Rater +
                     (1|Artifact), data=TxtOrg.ratings, REML=F)

mlm2_TxtOrg = lmer(Rating ~ 1 + (  Rater  )  + (1|Artifact),
                   data=TxtOrg.ratings, REML=F)

BIC(lmer_TxtOrg)
```

```
## [1] 263.2972
```

```r
BIC(mlm2_TxtOrg)
```

```
## [1] 264.6753
```

```r
BIC(mlm1_TxtOrg)
```

```
## [1] 299.7766
```

```r
summary(mlm2_TxtOrg)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: Rating ~ 1 + (Rater) + (1 | Artifact)
##    Data: TxtOrg.ratings
##
```

```
##      AIC      BIC   logLik deviance df.resid
##    250.9    264.7   -120.4    240.9      112
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3871 -0.5876  0.3244  0.5639  2.1462
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07498  0.2738
##  Residual             0.38752  0.6225
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.7590     0.1083  25.473
## Rater2       -0.1779     0.1493  -1.192
## Rater3       -0.3225     0.1493  -2.160
##
## Correlation of Fixed Effects:
##        (Intr) Rater2
## Rater2 -0.689
## Rater3 -0.689  0.500
```

```
summary(lmer_TxtOrg)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: TxtOrg.ratings
##
## REML criterion at convergence: 249
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3638 -0.7641  0.3836  0.5278  2.4094
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09145  0.3024
##  Residual             0.39503  0.6285
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.59144    0.06764   38.31
```

```
anova(lmer_TxtOrg, mlm2_TxtOrg)
```

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

```
## Data: TxtOrg.ratings
## Models:
```

```
## lmer_TxtOrg: Rating ~ 1 + (1 | Artifact)
## mlm2_TxtOrg: Rating ~ 1 + (Rater) + (1 | Artifact)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer_TxtOrg    3 251.45 259.74 -122.73   245.45
## mlm2_TxtOrg    5 250.86 264.68 -120.43   240.86 4.5892  2     0.1008
```

```
icc_txt2 = 0.09145 / (0.09145 + 0.39503)
icc_txt3 = 0.04436 / (0.04436 + 0.39256)
```

Similarly for text org we do not find any meaningful effects.

```
library(knitr)
ICC.vec.null = c(icc_crit, icc_init, icc_interp,icc_rsrch, icc_sel,
                 icc_txt, icc_vis)

f = rbind(Rubric.names , round(ICC.vec.null,2), round(ICC.vec.small,2),
          round(ICC.Vec.all,2))
rownames(f) = c('Rubric', 'Intercept', 'BIC (common)', 'BIC (all)')
kable(f)
```

	(Intercept)	(Intercept)	(Intercept)	(Intercept)	(Intercept)	(Intercept)	(Intercept)
Rubric	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg
Intercept	0.67	0.69	0.22	0.21	0.47	0.19	0.66
BIC (common)	0.57	0.49	0.23	0.19	0.52	0.14	0.59
BIC (all)	0.64	0.69	0.2	0.21	0.44	0.19	0.67

In the above table we aggregate all the ICCs, where BIC_all comes from variable selected highest BIC fixed effect model on the full dataset, BIC_common comes from the best BIC fixed effect model we found for the common data, and Intercept refers to a random intercept model. All of these also have random intercept effects for artifact. Most of these ICCs are very similar across all rubrics, suggesting that in most cases adding fixed effects for rater, semester, sex, repeated, etc. does not impact agreement between raters. However, The BIC_common values are different from the BIC_all values for a few rubrics, suggesting the semester and rater terms could be meaningful in determining ratings or rater agreement, and that the relationship between these terms differs across rubric.

```


tall[which(tall$Sex == F),"Sex"] = 'F'
mlm_all = lme4::lmer(Rating ~ 1 + (Sex   + Repeated   )* Rater * Rubric *
                       Semester + (0 + Rubric|Artifact), data=tall, REML=F)
```

Aggregate Model

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

```
mlm_all2 = lme4::lmer(Rating ~ 1 + (Sex + Rater  + Repeated + Semester)+
                        Rubric + (0 + Rubric|Artifact), data=tall, REML=F)

BIC(mlm_all)
```

```
## [1] 2304.711
```

```
BIC(mlm_all2)
```

```
## [1] 1671.394
```

```
#summary(mlm_all)
library(LMERConvenienceFunctions)
mlm_all_small= fitLMER.fnc(mlm_all,ran.effects=c("(1|Semester)","(1|Repeated)",
                                                  "(1|Rater)","(Semester|Rater)",
                                                  "(1|Sex)", "(1|Rubric)",
                                                  "(1|Artifact)",
                                                  "(Rater|Artifact)"),
                           method="BIC", alpha = 0.05)
```

```
## ======================================================
## ===              backfitting fixed effects         ===
## ======================================================
## setting REML to FALSE
```

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

```
## processing model terms of interaction level 4
##   iteration 1
##     p-value for term "Sex:Rater:Rubric:Semester" = 0.9141 >= 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| = 0.00280283 (tol = 0.002, component 1)
```

```
##     BIC simple = 2229; BIC complex = 2305; decrease = -76 < 5
##     removing term
##   iteration 2
##     p-value for term "Repeated:Rater:Rubric:Semester" = 0.5629 >= 0.05
##     not part of higher-order interaction
```

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

```
##     BIC simple = 2160; BIC complex = 2229; decrease = -70 < 5
##     removing term
## processing model terms of interaction level 3
##   iteration 3
##     p-value for term "Sex:Rater:Semester" = 0.6147 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 2147; BIC complex = 2160; decrease = -13 < 5
##     removing term
##   iteration 4
##     p-value for term "Repeated:Rubric:Semester" = 0.6237 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 2111; BIC complex = 2147; decrease = -36 < 5
##     removing term
##   iteration 5
##     p-value for term "Rater:Rubric:Semester" = 0.611 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 2040; BIC complex = 2111; decrease = -71 < 5
##     removing term
##   iteration 6
##     p-value for term "Sex:Rater:Rubric" = 0.55 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 1970; BIC complex = 2040; decrease = -70 < 5
##     removing term
##   iteration 7
##     p-value for term "Repeated:Rater:Semester" = 0.3065 >= 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| = 0.00233598 (tol = 0.002, component 1)
```

```
##     BIC simple = 1958; BIC complex = 1970; decrease = -11 < 5
##     removing term
##   iteration 8
##     p-value for term "Repeated:Rater:Rubric" = 0.2367 >= 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| = 0.00684271 (tol = 0.002, component 1)
```

```
##     BIC simple = 1892; BIC complex = 1958; decrease = -66 < 5
##     removing term
##   iteration 9
##     p-value for term "Sex:Rubric:Semester" = 0.2407 >= 0.05
##     not part of higher-order interaction
```

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

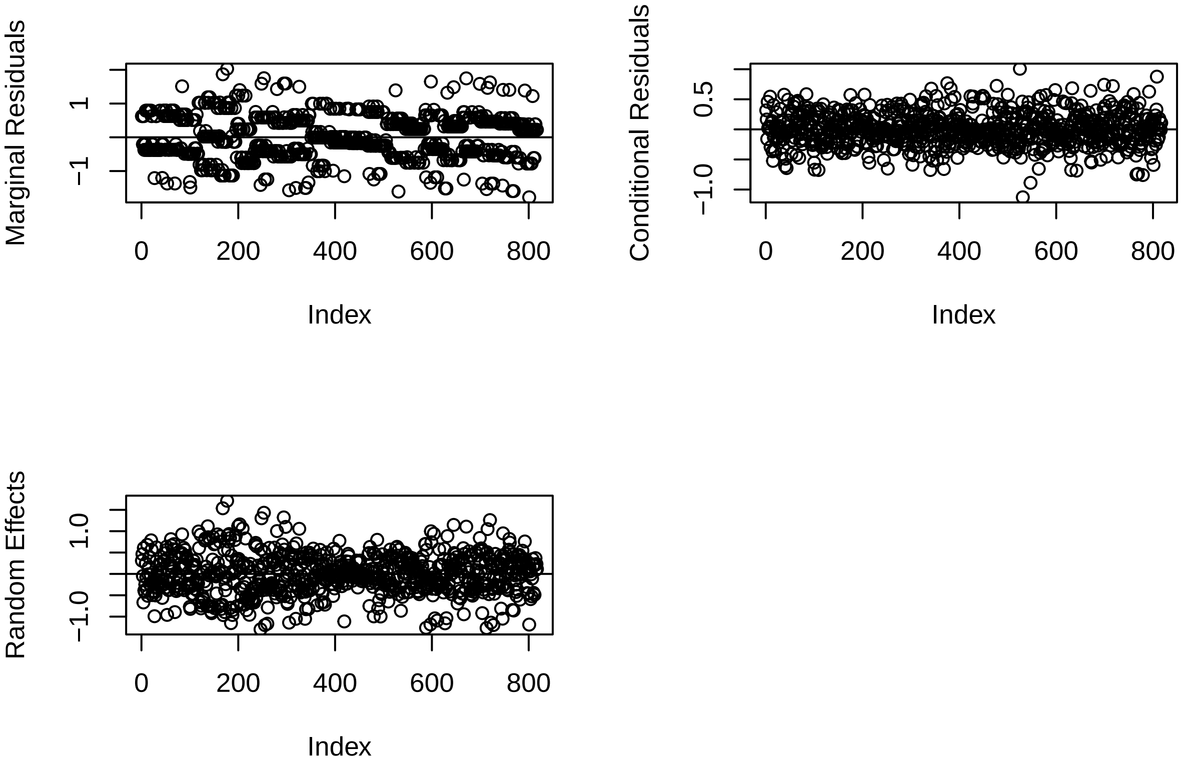
```
##     BIC simple = 1860; BIC complex = 1892; decrease = -33 < 5
##     removing term
## processing model terms of interaction level 2
##   iteration 10
##     p-value for term "Repeated:Semester" = 0.8989 >= 0.05
##     not part of higher-order interaction
```

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

```
##     BIC simple = 1854; BIC complex = 1860; decrease = -6 < 5
##     removing term
##   iteration 11
##     p-value for term "Rater:Semester" = 0.4391 >= 0.05
##     not part of higher-order interaction
```

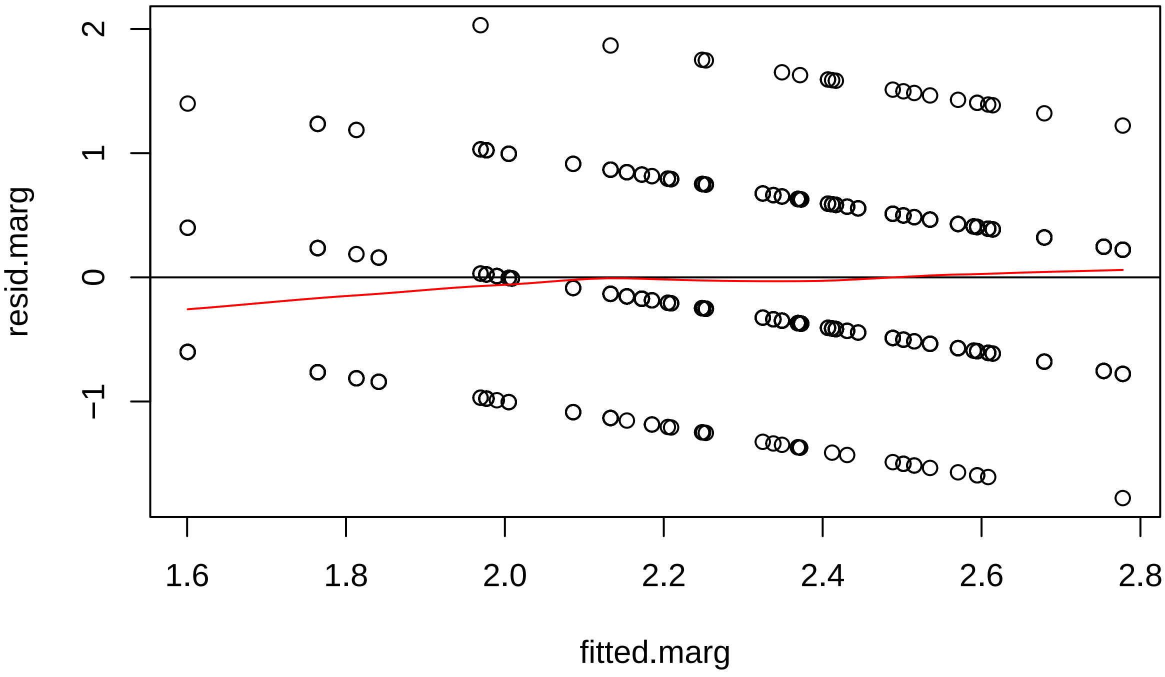
```
## boundary (singular) fit: see ?isSingular
##     BIC simple = 1842; BIC complex = 1854; decrease = -12 < 5
##     removing term
##   iteration 12
##     p-value for term "Repeated:Rubric" = 0.3707 >= 0.05
##     not part of higher-order interaction

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0109581 (tol = 0.002, component 1)

##     BIC simple = 1807; BIC complex = 1842; decrease = -35 < 5
##     removing term
##   iteration 13
##     p-value for term "Sex:Rubric" = 0.2967 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     BIC simple = 1773; BIC complex = 1807; decrease = -35 < 5
##     removing term
##   iteration 14
##     p-value for term "Repeated:Rater" = 0.269 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 1761; BIC complex = 1773; decrease = -11 < 5
##     removing term
##   iteration 15
##     p-value for term "Sex:Semester" = 0.2317 >= 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| = 0.00379329 (tol = 0.002, component 1)

##     BIC simple = 1756; BIC complex = 1761; decrease = -5 < 5
##     removing term
##   iteration 16
##     p-value for term "Rubric:Semester" = 0.0556 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     BIC simple = 1727; BIC complex = 1756; decrease = -28 < 5
##     removing term
##   iteration 17
##     p-value for term "Sex:Rater" = 0.1016 >= 0.05
##     not part of higher-order interaction

## boundary (singular) fit: see ?isSingular

##     BIC simple = 1718; BIC complex = 1727; decrease = -9 < 5
##     removing term
## processing model terms of interaction level 1
##   iteration 18
##     p-value for term "Sex" = 0.6841 >= 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| = 0.0219244 (tol = 0.002, component 1)

##     BIC simple = 1712; BIC complex = 1718; decrease = -7 < 5
##     removing term
##   iteration 19
##     p-value for term "Repeated" = 0.111 >= 0.05
##     not part of higher-order interaction
##     BIC simple = 1706; BIC complex = 1712; decrease = -6 < 5
##     removing term
## pruning random effects structure ...
##   nothing to prune
## ======================================================
## ===            forwardfitting random effects       ===
## ======================================================
## evaluating addition of (1|Semester) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 1
##  not adding (1|Semester) to model
## evaluating addition of (1|Repeated) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9988968
##  not adding (1|Repeated) to model
## evaluating addition of (1|Rater) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 1
##  not adding (1|Rater) to model
## evaluating addition of (Semester|Rater) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 1
##  not adding (Semester|Rater) to model
## evaluating addition of (1|Sex) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 1
##  not adding (1|Sex) to model
## evaluating addition of (1|Rubric) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 0.9991981
##  not adding (1|Rubric) to model
## evaluating addition of (1|Artifact) to model
## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 1
##  not adding (1|Artifact) to model
## evaluating addition of (Rater|Artifact) to model

## boundary (singular) fit: see ?isSingular

##  log-likelihood ratio test p-value = 2.218981e-09
##  adding (Rater|Artifact) to model
## ========================================================
## ===            re-backfitting fixed effects          ===
## ========================================================
## setting REML to FALSE

## boundary (singular) fit: see ?isSingular

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## Warning in pf(anova.table[term, "F value"], anova.table[term, "npar"],
## nrow(model@frame) - : NaNs produced

## 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
## log file is /var/folders/g4/7xrypdv52yx9nr_1mv034x_00000gn/T//RtmpuoVhDD/fitLMER_log_Fri_Dec_10_15-25
```

```
par(mfrow=c(2,2))
plot(r.marg(mlm_all_small),xlab="Index",ylab="Marginal Residuals")
abline(0,0)
plot(r.cond(mlm_all_small),xlab="Index",ylab="Conditional Residuals")
abline(0,0)
plot(r.reff(mlm_all_small),xlab="Index",ylab="Random Effects")
abline(0,0)
```

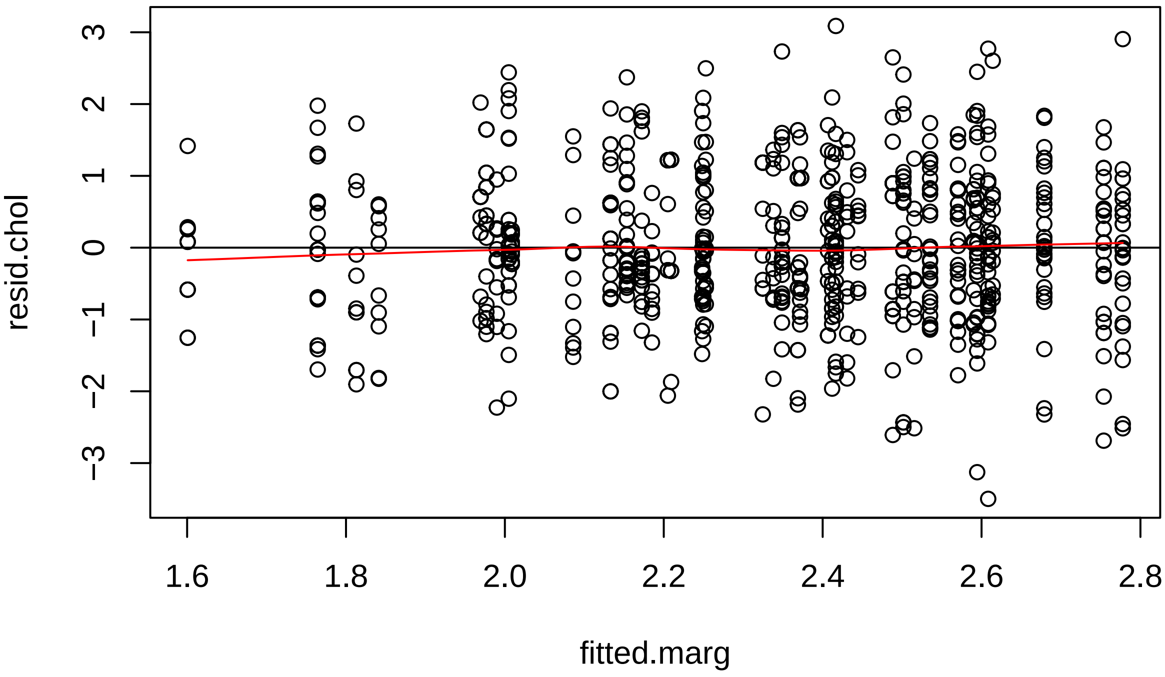


Diagnostics

```
fitted.marg = yhat.marg(mlm_all_small)
resid.marg = r.marg(mlm_all_small)
resid.chol = r.chol(mlm_all_small)
plot(fitted.marg,resid.marg)
abline(h=0)
lines(loess.smooth(fitted.marg,resid.marg),col="red")
```



```
plot(fitted.marg,resid.chol)
abline(h=0)
lines(loess.smooth(fitted.marg,resid.chol),col="red")
```



```
summary(mlm_all_small)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + Rubric + Semester + (0 + Rubric | Artifact) +
##     (Rater | Artifact) + Rater:Rubric
##    Data: tall
##
## REML criterion at convergence: 1380.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.07974 -0.46668 -0.03091  0.45361  2.74806
##
## Random effects:
##  Groups     Name            Variance Std.Dev. Corr
##  Artifact   RubricCritDes   0.49401  0.7029
##             RubricInitEDA   0.31089  0.5576    0.32
##             RubricInterpRes 0.09991  0.3161    0.15  0.67
##             RubricRsrchQ    0.17706  0.4208    0.50  0.19  0.54
##             RubricSelMeth   0.03795  0.1948    0.16  0.22  0.38 -0.23
##             RubricTxtOrg    0.24203  0.4920    0.27  0.43  0.35  0.30  0.19
##             RubricVisOrg    0.22683  0.4763    0.18  0.50  0.44  0.27 -0.16
##  Artifact.1 (Intercept)     0.01406  0.1186
##             Rater2          0.16135  0.4017   -0.64
##             Rater3          0.08992  0.2999    0.04  0.75
##  Residual                   0.13436  0.3665
##
##
##
##
##
##
##
##   0.53
##
##
##
##
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)             1.76438    0.11383  15.500
## Rater2                  0.36865    0.13914   2.649
## Rater3                  0.21240    0.12966   1.638
## RubricInitEDA           0.73727    0.12943   5.696
## RubricInterpRes         0.98940    0.12713   7.783
## RubricRsrchQ            0.72392    0.11748   6.162
## RubricSelMeth           0.40803    0.12409   3.288
## RubricTxtOrg            1.01340    0.12949   7.826
## RubricVisOrg            0.65224    0.13288   4.909
## SemesterS19            -0.16368    0.07713  -2.122
## Rater2:RubricInitEDA   -0.29989    0.15575  -1.925
## Rater3:RubricInitEDA   -0.30213    0.15541  -1.944
## Rater2:RubricInterpRes -0.51408    0.15309  -3.358
## Rater3:RubricInterpRes -0.71656    0.15265  -4.694
## Rater2:RubricRsrchQ    -0.48813    0.14687  -3.324
## Rater3:RubricRsrchQ    -0.32782    0.14627  -2.241
## Rater2:RubricSelMeth   -0.38748    0.14989  -2.585
## Rater3:RubricSelMeth   -0.37990    0.14867  -2.555
## Rater2:RubricTxtOrg    -0.55192    0.15611  -3.536
## Rater3:RubricTxtOrg    -0.45497    0.15576  -2.921
## Rater2:RubricVisOrg    -0.10627    0.15817  -0.672
## Rater3:RubricVisOrg    -0.28021    0.15782  -1.776
```

```
##
## 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
```

```
BIC(mlm_all_small)
```

```
## [1] 1762.999
```

```
anova( mlm_all_small, mlm_all2)
```

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

```
## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.
```

```
## Data: tall
## Models:
## mlm_all2: Rating ~ 1 + (Sex + Rater + Repeated + Semester) + Rubric + (0 + Rubric | Artifact)
## mlm_all_small: Rating ~ Rater + Rubric + Semester + (0 + Rubric | Artifact) + (Rater | Artifact) + Ra
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mlm_all2        41 1478.5 1671.4 -698.23   1396.5
## mlm_all_small   57 1425.9 1694.1 -655.94   1311.9 84.574 16  2.468e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kable(round(summary(mlm_all_small)$coef, 2))
```

	Estimate	Std. Error	t value
(Intercept)	1.76	0.11	15.50
Rater2	0.37	0.14	2.65
Rater3	0.21	0.13	1.64
RubricInitEDA	0.74	0.13	5.70
RubricInterpRes	0.99	0.13	7.78
RubricRsrchQ	0.72	0.12	6.16
RubricSelMeth	0.41	0.12	3.29
RubricTxtOrg	1.01	0.13	7.83
RubricVisOrg	0.65	0.13	4.91
SemesterS19	-0.16	0.08	-2.12
Rater2:RubricInitEDA	-0.30	0.16	-1.93

	Estimate	Std. Error	t value
Rater3:RubricInitEDA	-0.30	0.16	-1.94
Rater2:RubricInterpRes	-0.51	0.15	-3.36
Rater3:RubricInterpRes	-0.72	0.15	-4.69
Rater2:RubricRsrchQ	-0.49	0.15	-3.32
Rater3:RubricRsrchQ	-0.33	0.15	-2.24
Rater2:RubricSelMeth	-0.39	0.15	-2.59
Rater3:RubricSelMeth	-0.38	0.15	-2.56
Rater2:RubricTxtOrg	-0.55	0.16	-3.54
Rater3:RubricTxtOrg	-0.45	0.16	-2.92
Rater2:RubricVisOrg	-0.11	0.16	-0.67
Rater3:RubricVisOrg	-0.28	0.16	-1.78

Finally, we fit a combined model on all the data using every explanatory variable and interactions with rater, rubric, and semester. We perform automatic variable selection using BIC criterion and arrive at a final model with fixed effects for rater, rubric, semester, and interactions between rater and rubric. The model also has random uncorrelated slopes for rater and rubric grouped by artifact. In summary, this model means that there rater, rubric, and semester are all important in determining a student's rating, and that the relationship between rating and rater varies across rubrics. The random effects also tell us that there are differences in the relationship between rater and rating as well as the relationship between rubric and rating that cannot be fully captured by the fixed effects. One possible explanation for this is the differing degrees of variation across rubric in the random effects. For example, the tau^2 for CritDes is much higher than the others, indicating more variation in scores for this rubric item across all artifacts. The tau^2 for rater 2 is also higher than that of rater 3, possibly indicating more variation in the scores of rater 2 across artifact.

From the residual plots, we can see that there is only slight nonconstant variance, likely due to sample size.

Research Question 4

Are there any other interesting properties of the data that were not addressed in the previous three questions?

All the relevant technical details for Research Question 4 are contained in the EDA for Q1 and the model diagnostics from Q3. In lieu of a technical exposition for this section, we present the similar text as shown in the paper that discusses our general findings:

One interesting thing in the data is the fact that there were a few missing observations that we simply imputed because we gathered that the distribution would not be affected much depending on our imputation. However, it could be possible that the missing data is systematic, e.g. the missing sex information could be due to a reporting difference that might be representative of an entirely new factor class for this variable. In future work this should be investigated further.

In terms of context based conclusions, we were able to identify key variables that had important explanatory relationships with rating and we were able to quantify these relationships. Perhaps the most important of these variables is the rater; we noticed differences in rater behavior (i.e. how they scored artifacts, like how rater 3 was a harsher grader than rater 1 from the EDA), and that these behaviors are different for different rubrics. We were also able to determine that ratings differ across semesters, but the relationship between ratings and rubric and rater are all fairly constant across semesters for most rubrics. The notion that raters behaved similarly across semesters for multiple rubrics is important for the Dean to know, as this is evidence for the fairness of the experimental grading process across different iterations of the seminar. However, because this is only true for certain individual rubrics, it is possible that there does exist some variation in how raters behave or how certain rubrics are percieved across semesters. Also, the raters themselves behave differently, so it might be a good idea to add more raters in the future to prevent this variability from becoming apparent in student grades.

Another interesting point is that there is a discrepancy between our EDA findings and our model coefficients. Our EDA demonstrated that rater 3 was the harshest, but the model assigned them a positive coefficient, indicating they tend to give out higher scores than rater 1. However, when accounting for interactions, this is not the case. For example, the estimated decrease in rating from rater 1 to rater 3, when considering interactions and specifying the rubric as InterpRes, is 0.21-0.51 = -0.3. All the interaction coefficients for rater 3 are negative and relatively large in magnitude, meaning that when considering these terms for each rubric the model does estimate rater 3 to be the harshest. This shows us that the random and fixed effects, as well as the interactions, are meaningful, and should be taken into consideration with the conclusions drawn from earlier visualizations.

We can also notice other interesting relationships from the EDA we did (See research question 1). In particular, we noticed that ratings do not change much with sex, but do change with semester. So, other factors besides rating are meaningful in this dataset.

Finally, we take another look at thhe diagnostic plots from section 3 to determine if the fit is adequate. Overall, the variance of the residuals looks fairly constant and centered around 0, indicating that our assumption of treating rating as a quantitative response was reasonable.