Assessment of the Rating for "General Education" program for Undergraduates

Naijia Liu
naijial@andrew.cmu.edu

December 8, 2021

Abstract

The Dietrich College at Carnegie Mellon University is interested in evaluating the student performance in their new general education program. This study aims to analyze the recent experimentation performed by the college to see the associations and distributions between the ratings. We examined data on 91 project papers that were randomly sampled from a fall and spring section of freshman statistics that was rated by three raters using seven different rubrics. From our analysis, we used visual plot to identify the distribution of ratings for each rubrics and given by each rater, and the intraclass correlation was used as the measurement of agreement between the raters, and multiple mixed linear models were fitted to test how the various factors are related to the ratings. We find that the ratings are not entirely indistinguishable among different raters and rubrics, and for the rubric *RsrchQ*, *InitEDA*, *InterpRes* and *TxtOrg*, raters are inconsistent with one another in how they rate. In the final multi-level model consists of fixed effects, random effects and interaction terms between the variables *Rater*, *Semester* and *Rubric*. Moreover, it would be worthwhile to consider the fall, spring semester data imbalance and larger sample size when repeating this experiment in the future.

1 Introduction

Dietrich College at Carnegie Mellon University is in the process of implementing a new "General Education" program for undergraduates. This program specifies a set of courses and experiences that all undergraduates must take, and in order to determine whether the new program is successful, the college hopes to rate student work performed in each of the "Gen Ed" courses each year.

In this paper, we have been asked by the associate dean in charge of this experiment to assess the rating work in Freshman Statistics, which uses 3 raters from across the college. To be more specific, we will explore the influence of different raters and different rubrics assigned by artifacts on grading work. Moreover, the relationship with various factors and rating in this experiment will be probed into.

In addition to answering the main question posed above, we will address the following questions:

- Is the distribution of ratings for each rubrics pretty much indistinguishable from the other rubrics, or are there rubrics that tend to get especially high or low ratings? Is the distribution of ratings given by each rater pretty much indistinguishable from the other raters, or are there raters that tend to give especially high or low ratings?
- For each rubric, do the raters generally agree on their scores? If not, is there one rater who disagrees with the others? Or do they all disagree?
- More generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) related to the ratings? Do the factors interact in any interesting ways?
- Is there anything else interesting to say about this data?

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 eval- uates to what extent a study design convincingly answer that question.
InitEDA	Initial EDA	Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis.
SelMeth	Select Method(s)	Given a data set and a research question, the student selects appropriate method(s) to analyze the data.
InterpRes	Interpret Results	The student appropriately interprets the results of the selected method(s).
VisOrg	Visual Organization	The student communicates in an organized, coherent and effective fashion with visual elements (charts, graphs, tables, etc.).
TxtOrg	Text Organization	The student communicates in an organized, coherent and effective fashion with text elements (words, sentences, paragraphs, section and subsection titles, etc.).

Table 1: Rubrics for rating Freshman Statistics projects.

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

Table 2: Rating scale used for all rubrics.

2 Data

In a recent experiment, 91 project papers that are referred to as "artifacts" were randomly sampled from a Fall and Spring section of Freshman Statistics. Three raters from three different departments were asked to rate these artifacts on seven rubrics, as shown in Table 1. The rating scale for all rubrics is shown in Table 2. The raters did not know which class or which students produced the artifacts that they rated. Thirteen of the 91 artifacts were rated by all three raters; each of the remaining 78 artifacts were rated by only rater. The variables available for analysis are defined in Table 3.

3 Methods

The first research question focuses on the distributions of ratings between rubrics and raters, more specifically, to examine if there is any disparity in ratings by rubric or rater. We have looked into the distribution using the usual one-dimensional summary statistics, the table of counts and the histogram for each rubric and based on each rater (See Figure 1 and Figure 2). What we also considered is whether 13 artifacts are representative of the whole set of 91 artifacts, and the comparison is also performed for the subset of data that contains the 13 artifacts that were rated by all three raters to corroborate conclusions. Detailed analysis can be found in Appendix 1.

Then, to address the question about whether the raters generally agree on their scores for each rubric, and is there one rater who disagrees with the others or do they all disagree, the measurement of agreement among the raters we used is the intraclass correlation (ICC). We the subset of the data for just the 13 artifacts seen by all three raters, and fitted seven random-intercept models, one for each rubric, and calculate the seven ICC's. After that, by making a 2-way table of counts for the ratings of each pair of raters and on each rubric, we shall have the the percent exact agreement between the two raters from the percentage of observations on the main diagonal, which helps us to determine who is agreeing with whom on each rubric. We also did the ICC calculations on the full data set to see whether the seven ICC's for the full data set agree with the seven ICC's for the subset corresponding

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 Which semester the artifact came from
Sex	M or F	Sex or gender of student who created the artifact
RsrchQ	1, 2, 3 or 4	Rating on Research Question
CritDes	1, 2, 3 or 4	Rating on Critique Design
InitEDA	1, 2, 3 or 4	Rating on Initial EDA
SelMeth	1, 2, 3 or 4	Rating on Select Method(s)
InterpRes	1, 2, 3 or 4	Rating on Interpret Results
VisOrg	1, 2, 3 or 4	Rating on Visual Organization
TxtOrg	1, 2, 3 or 4	Rating on Text Organization
Artifact	(text labels)	Unique identifier for each artifact
Repeated	0 or 1	1 = this is one of the 13 artifacts seen by all 3 raters

Table 3: Definition of all variables.

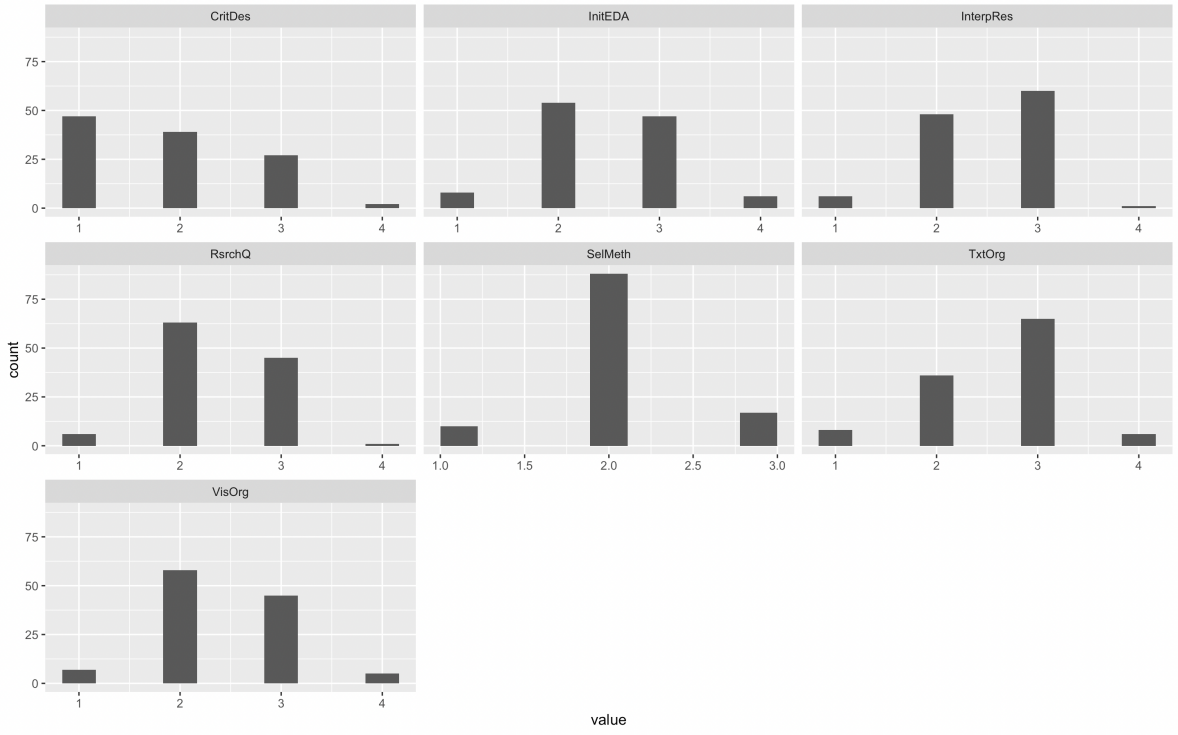


Figure 1: Distributions of Ratings across Rubrics.

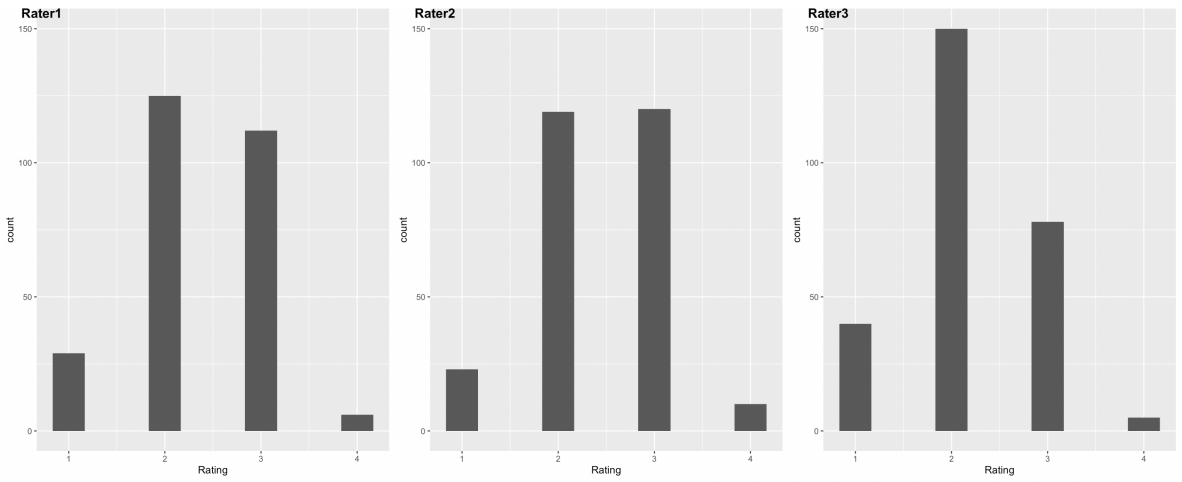


Figure 2: Distributions of Ratings across Raters.

Rubrics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.36	3.0	4	0.60
CritDes	1	1	2	1.86	2.5	4	0.84
InitEDA	1	2	2	2.44	3.0	4	0.70
SelMeth	1	2	2	2.06	2.0	3	0.48
InterpRes	1	2	3	2.49	3.0	4	0.61
VisOrg	1	2	2	2.42	3.0	4	0.68
TxtOrg	1	2	3	2.60	3.0	4	0.70

Table 4: Numeric Summary for Each Rubric based on the Full Data.

to the 13 artifacts that all three raters saw. Detailed analyses can be found in Appendix 2.

Next, we considered more generally, how are the various factors in this experiment (Rater, Semester, Sex, Repeated, Rubric) are related to the ratings, also in R, by fitting the linear mixed-effects models. First, using the artifact as the random grouping variable, we added fixed effects for rater, semester, sex and repeated to the random intercept models for the full data set. In particular, the ANOVA test with the AIC and BIC will be used to determine the significance of the added effects, and the backward elimination selection method will also be used to see which fixed effect should be added to the model as further evidence. Next, corresponding interaction terms will be added and tested using ANOVA and the lmer function.

Since each model considers only one rubric at a time, we switched to another data set and repeated the same process as above to explore interactions with rubric directly. Finally, the final model was chose, and was analyzed using its summary regression statistics. Detailed R analyses can be found in Appendix 3.

4 Results

4.1 Distribution of Ratings

For each rubrics, we assumed that ratings with values that are less than or equal to 2 for each rubrics are considered as low, and that ratings with values that are higher than 2 for each rubrics are considered as high.

Rubric *SelMeth* (Rating on Select Method(s)) tend to get especially low ratings. That is because the 3rd quantile for *SelMeth* is 2 (See Table 4), which is the lowest among all the rubrics. This means that at least 75% of artifacts get score lower than 2 for rubric *SelMeth*. And the max score for rubric *Selmeth* is 3, which is also lower than all the other rubrics. We can also see from the table of counts that 99 artifacts get score that is equal or lower than 2, which collides with our findings in the Figure 1.

Though, the percentage that artifacts scored in 1 of the rubric *CritDes* is the lowest among all

Rubrics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
Rater1	1	2	2	2.35	3	4	0.70
Rater2	1	2	2	2.43	3	4	0.70
Rater3	1	2	2	2.18	3	4	0.69

Table 5: Numeric Summary for Each Rater based on the Full Data.

the other rubrics and has largest number of rating 1, the total amount of artifacts scored less than or equal to 2 is lower than *SelMeth*. So, we are still apt to draw the conclusion that rubric *SelMeth* tend to get especially low ratings.

Rubric *InterpRes* (Rating on Interpret Results) and *TxtOrg* (Rating on Text Organization) tend to get especially high ratings. That is because the median for both rubrics is 3, which is higher than the others. This implies that at least 50% of artifacts get score higher than 3 for these two rubrics. The mean for *InterpRes* is 2.49 and the mean for *TxtOrg* is 2.60, which are the top two highest among all rubrics. We can also see from the histogram that over 60 of artifacts score higher or equal to 3, which corresponds to what we have found in the statistics summary table.

Across raters, we also assumed that ratings with values that are less than or equal to 2 for each rubrics are considered as low, and that ratings with values that are higher than 2 for each rubrics are considered as high. From the statistics summary table (See Table 5), the mean for all 3 raters are pretty similar, and the SD for all 3 raters are pretty similar too. Thus, the distribution of ratings given by each rater is pretty much indistinguishable from the other raters.

However, from the histogram (See Figure 2) and the corresponding table of counts above, it seems that rater3 tend to give lower ratings than other 2 raters. That is because rater 3 have the highest percent of scoring 2, and it has the highest total number of ratings that are equal or lower than 2 among all 3 raters. Though, the distribution of ratings given by rater 1 and rater 2 is pretty much indistinguishable from each other.

Using the same method as above, we can also draw a conclusion that these 13 artifacts are representative of the whole set of 91 artifacts, because the histograms of the ratings for each rubric and rater show identical patterns to the summary statistics for the full and subset of data. And from the statistics summary, the 3rd quantile, the mean and the standard deviation, minimum value, 1st quantile and the median are identical in both data set (See Appendix 1, pg. 12). However, the only small deviations are the lower average rating for rater three and rater three did not give a rating of four in the subset. However, it is reassuring that the ratings for the subset of data is consistent between raters for other parts of the analysis.

4.2 Agreement Among the Raters

First, the intra-cluster correlation (ICC) is used to measure agreement by determining the correlation between any two rater's ratings on the same artifact. First we examine the 13 "common" artifacts that all 3 raters saw. As we consider the value of an intra-class correlation that is less than 0.50 as poor reliability, and the value that is between 0.5 and 0.75 as moderate reliability. Thus, we could say that, based on the rules of thumb for interpreting ICC, the rubric *RsrchQ*, *InitEDA*, *InterpRes* and *TxtOrg* can be rated with "poor" reliability by different raters, which means the raters are inconsistent with one another in how they rate; while, *CritDes*, *SelMeth* and *VisOrg* can be can be rated with "good" reliability by different raters, which means the raters do not agree with their scores for all rubrics. The higher ICC of the rubric is, the more raters agree (see Table 6).

Then, the percent of the exact agreement of each pair of raters on each rubric will be calculated to tell which raters might be contributing to disagreement.

The percentage of observations for rubric *InitEDA* with Rater1 and Rater3 is the lowest among all 3 pairs, and the higher proportion for Rater2 and Rater3 indicates that the Rater1 has significantly different ratings. The percentage of observations for rubric *InterpRes* are pretty similar among all 3 pairs, which indicates that all three raters have significantly different ratings for rubric *InterpRes*. The percentage of observations for rubric *RsrchQ* with Rater1 and Rater2 is the lowest among all 3 pairs, and the higher proportion for Rater1 and Rater3 indicates that the Rater2 has significantly different ratings. The percentage of observations for rubric *TxtOrg* with Rater2 and Rater3 is the lowest among all 3 pairs, and the higher proportion for Rater2 and Rater1 indicates that the Rater3 has significantly

Rubric	ICC.alldata	ICC.common	a12	a23	a13
CritDes	0.67	0.57	0.54	0.69	0.62
InitEDA	0.69	0.49	0.69	0.85	0.54
InterpRes	0.22	0.23	0.62	0.62	0.54
RsrchQ	0.21	0.19	0.38	0.54	0.77
SelMeth	0.47	0.52	0.92	0.69	0.62
TxtOrg	0.19	0.14	0.69	0.54	0.62
VisOrg	0.66	0.59	0.54	0.77	0.77

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

	CritDes	InitEDA	InterpRes	RsrchQ	SelMeth	TxtOrg	VisOrg
(Intercept)	–	–	–	–	–	–	–
Repeated	–	–	–	–	–	–	–
SemesterS19	–	–	–	–	-0.35860	–	–
Sex	–	–	–	–	–	–	–
Rater1	1.6863	–	2.70421	–	2.25037	–	2.37794
Rater2	2.1129	–	2.58574	–	2.22653	–	2.64891
Rater3	1.8908	–	2.13918	–	2.03316	–	2.28355
σ^2	0.2473	0.1655	0.25250	0.27825	0.10842	0.39573	0.1467

Table 7: Estimated coefficients for models of each rubric.

different ratings (see Table 6).

For the ICC calculations on the full data set, we do not have the percent exact agreement calculations with the subset, that is because the other 78 artifacts are only graded by only one grader, and there is no way to compare the rating between any 2 raters on the same rubric. To conclude, the raters do not generally agree on all their scores and each rater has a particular rubric where they disagreed with the other two raters.

4.3 Relationship between Ratings and Various Factors

More generally, to analyze how are the various factors in this experiment, including rater, semester, sex, repeated and rubric related to the ratings. First, we will try to add fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw. Using backward elimination variable selection (Appendix 3, pg. 18), we were able to find out that for all 7 models for all 7 rubrics, which can be written in the format of model 1.

$$Rating \sim (1|Artifact) \tag{1}$$

Each rubric's rating on each Artifact differs from what we would expect by a small random effect that depends on the Artifact. None of the fixed effect variables were retained, and there was no need to check for any interaction terms or additional random effects.

Similarly, we repeated the variable selection on the 7 rubric-specific models for the full data set. There are some differences among the models: For *InitEDA*, *RsrchQ* and *TxtOrg*, the models are just the simple random-intercept models, However, for the other four, the models are a little more complex. So, for these four models, we examined each of these 4 models to see if the fixed effects make sense to us, and if there are any interactions or additional random effects to consider using ANOVA t-tests (Appendix 3, pg. 18). The table 7 below is the result we got, and the results can be interpreted as follow.

- *SelMeth*: Considering the same rater in the same semester (ex. Semester Spring 19), the rating for rubric *SelMeth* is distinguishable in different artifacts. For the same artifact in the same semester, rater 3 gives the rating that is 0.19337 lower than rater 2, and 0.217 than rater 1. For the same artifact rated by the same rater, the rating for rubric 'SelMeth' in Semester Spring 19 is 0.359 lower than that in Semester Fall 19.

- *CritDes*: Considering the same rater, the rating for rubric is distinguishable in different artifacts, and for the same artifact, rater 2 gives the rating that is 0.427 higher than rater 1, and 0.222 higher than rater 3 for rubric *CritDes*.
- *InterpRes*: Considering the same rater, the rating for rubric is distinguishable in different artifacts, and for the same artifact, and rater 1 gives the rating that is 0.118 higher than rater 2, and 0.565 higher than rater 3 for rubric *InterpRes*.
- *VisOrg*: Considering the same rater, the rating for rubric is distinguishable in different artifacts, and for the same artifact, rater 2 gives the rating that is 0.271 higher than rater 1, and 0.365 higher than rater 3.
- *TxtOrg*: The rating for rubric is distinguishable in different artifacts.
- *RsrchQ*: The rating for rubric is distinguishable in different artifacts.
- *InitEDA*: The rating for rubric is distinguishable in different artifacts.

Finally, we will start trying to add fixed effects to the "combined" model 2 without having to fit 7 separate models using all the data. Through backwards elimination, fixed effect variables in the final model we selected. Next, we attempted to include all possible interaction terms using the selected fixed effect variables, then carried out backward elimination again to select the best subset of interaction terms (Appendix 3, pg.29).

From here, we were then able to use the AIC, BIC values as well as the likelihood ratio tests to compare models that included all the interaction terms, the subset of interaction terms after backward elimination, and no interaction terms at all. Finally, the second model 3 with selected subset of interaction terms is selected, and the coefficient summary of the chosen model can be seen in Table 8.

$$Rating \sim 1 + (0 + Rubric|Artifact) \tag{2}$$

$$\begin{aligned} Rating =& \beta_0 + \beta_1 * Rater2 + \beta_2 * Rater3 \\ &+ \beta_3 * SemesterS19 + \beta_4 * RubricInitEDA \\ &+ \beta_5 * RubricInterpRes + \beta_6 * RubricRsrchQ \\ &+ \beta_7 * RubricSelMeth + \beta_8 * RubricTxtOrg \\ &+ \beta_9 * RubricVisOrg + \beta_10 * Rater2 : RubricInitEDA \\ &+ \beta_{11} * Rater3 : RubricInitEDA + \beta_{12} * Rater2 : RubricInterpRes \\ &+ \beta_{13} * Rater3 : RubricInterpRes + \beta_{14} * Rater2 : RubricRsrchQ \\ &+ \beta_{15} * Rater3 : RubricRsrchQ + \beta_{16} * Rater2 : RubricSelMeth \\ &+ \beta_{17} * Rater3 : RubricSelMeth + \beta_{18} * Rater2 : RubricTxtOrg \\ &+ \beta_{19} * Rater3 : RubricTxtOrg + \beta_{20} * Rater2 : RubricVisOrg \\ &+ \beta_{21} * Rater3 : RubricVisOrg + (0 + Rubric|Artifact) \\ &+ (0 + Rater|Artifact) \end{aligned} \tag{3}$$

In our final model, we expect that rubric scores depend on artifact, and the artifacts will not be all of equal quality on each rubric. Average scores vary from rubric to rubric, and it also varies a bit from one artifact to the next, by a small random effect that depends on artifact. In all of this, we can say that each rater's rating on each artifact differs from what we would expect (from the fixed effects alone) by a small random effect that depends on the artifact due to the interaction of rater and artifact, which suggests that the raters are not interpreting the evidence in the artifacts in the same way. Each rater also uses each rubric in a way that is not like, or even parallel to, other rater's rubric usage, which can be proved in the facets plot (See Figure 3).

It does look as if the 3 raters have different ways of scoring the 7 rubrics, so the interaction we found in final model makes sense. Among all 3 raters, for rubric *InitEDA* and *VisOrg*, rater 2 tends to give the highest score, while rater 1 tends to give the highest score for other 5 rubrics. For example, for rubric *InitEDA*, given all the other variables are the same, rater 2 rates 0.067 higher than rater1

	Estimate	Std. Error
β_0 :(Intercept)	1.7575357	0.11402967
β_1 :Rater2	0.3660743	0.13917859
β_2 :Rater3	0.1959298	0.12965892
β_3 :SemesterS19	-0.1591747	0.07647292
β_3 :RubricInitEDA	0.7395208	0.12995961
β_4 :RubricInterpRes	0.9915188	0.12770181
β_5 :RubricRsrchQ	0.7262014	0.11791907
β_6 :RubricSelMeth	0.4107115	0.12469405
β_7 :RubricTxtOrg	1.0157913	0.12999164
β_8 :RubricVisOrg	0.6542375	0.13353097
β_9 :Rater2:RubricInitEDA	-0.2998406	0.15609130
β_{10} :Rater3:RubricInitEDA	-0.2947790	0.15635257
β_{11} :Rater2:RubricInterpRes	-0.5132331	0.15348295
β_{12} :Rater3:RubricInterpRes	-0.7148403	0.15363779
β_{13} :Rater2:RubricRsrchQ	-0.4874343	0.14721456
β_{14} :Rater3:RubricRsrchQ	-0.3224062	0.14725825
β_{15} :Rater2:RubricSelMeth	-0.3864167	0.15030393
β_{16} :Rater3:RubricSelMeth	-0.3871985	0.14960917
β_{17} :Rater2:RubricTxtOrg	-0.5510611	0.15645949
β_{18} :Rater3:RubricTxtOrg	-0.4449033	0.15673034
β_{19} :Rater2:RubricVisOrg	-0.1048823	0.15861238
β_{20} :Rater3:RubricVisOrg	-0.2751871	0.15885035

Table 8: Estimated coefficients for final model.

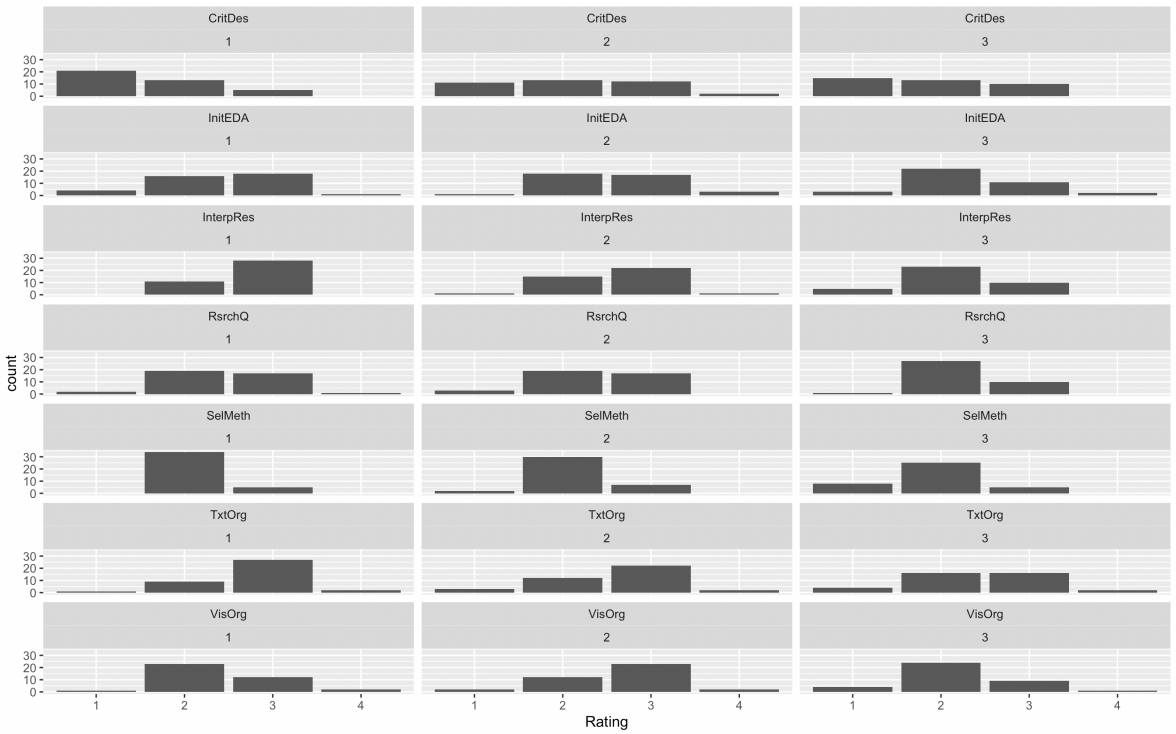


Figure 3: Facet Plot for 3 Raters across all Rubrics.

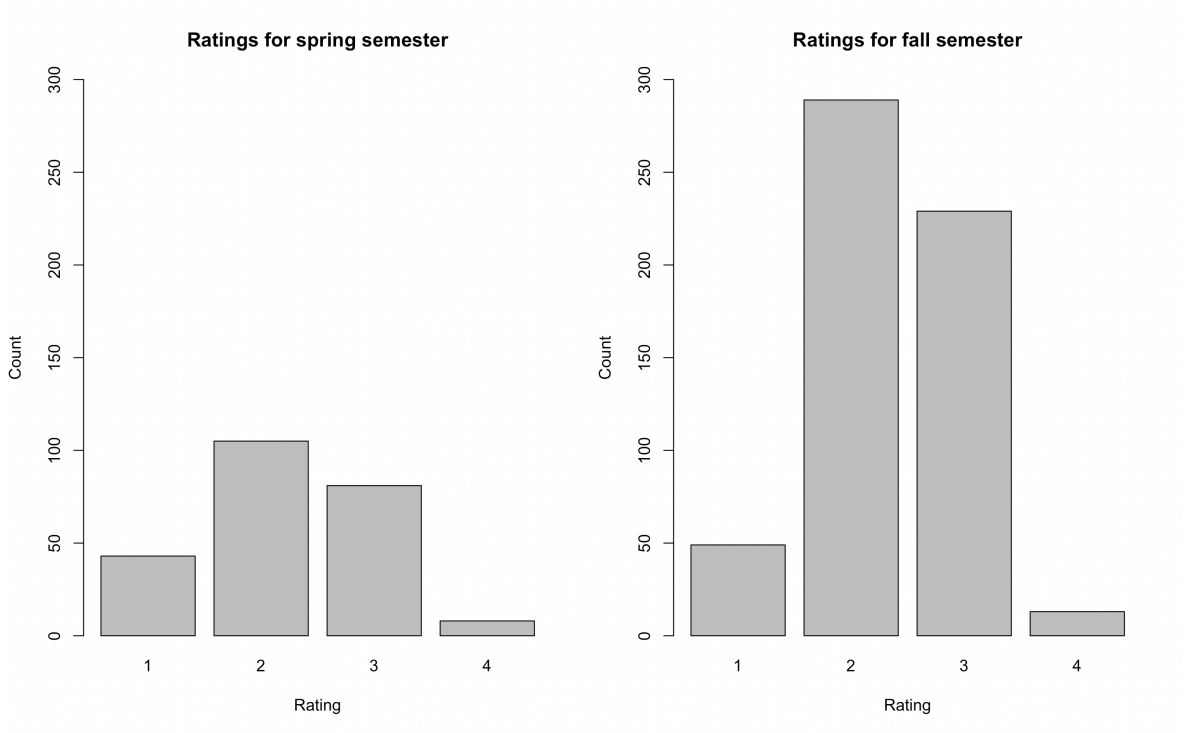


Figure 4: Proportion of ratings in the fall and spring semesters.

in semester S19. Among all rubrics, rater 1 tends to give the lowest score for rubric *SelMeth*, rater 2 tends to give the lowest score for rubric *SelMeth*, rater 3 tends to give the lowest score for rubric *CritDes*. For example, for rater 1, given all the other variables are the same, he/she will rate 0.605 less for rubric *SelMeth* than rubric *TxtOrg* in semester S19.

As Figure 3 suggests that the raters are not all interpreting the rubrics in the same way. These interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more similar to each other. Moreover, artifacts in semester S19 tend to get lower ratings than semester F19. For example, for any rubric rated by the same rater, given all the other variables are the same, artifacts in semester S19 will have grades 0.159 less than artifacts in semester F19.

4.4 Anything Else Interesting about Data

In the previous research question, as we have mentioned that artifacts in semester S19 will have grades less than artifacts in semester F19. Anticipating future questions regarding this result, the distributions of counts for rating of two semesters were further analyzed to investigate possible explanations. The raw counts of the artifacts per semester showed that more artifacts were sampled from the fall semester than the spring, and the distribution of the ratings shows that most artifacts received a score of two in both semesters as shown in Figure 4.

This finding is important since the inclusion of the semester variable in the final model, and recall that the coefficient for the semester variable is negative, which aligns with the results in Figure 4 since the proportion of artifacts received a one is larger in spring. That is to say, the artifacts from the spring received lower scores than the artifacts in the fall may not due to random variation in selection.

In the end of our analysis, we draw dour residual plots for the model we selected, all of which prove that it is a great model for the data set. What we also did is that we recalculated the ICC's for the new model, and compare them with the earlier ICC's, and the result is that ICC's from these models do not agree with our earlier ICC's (Appendix 4, pg.33). Though the difference is relatively small for rubric *VisOrg*, *InitEDA* and *CritDes*, while the difference is relatively large for rubric *TxtOrg*, *InterpRes*, *RsrchQ* and *SelMeth*, suggesting further investigation.

5 Discussion

From the histogram, numeric summary (See Figure 1) and the table of counts, we can tell that the distribution of ratings for each rubrics is pretty much indistinguishable from the other rubrics. Rubric *SelMeth* (Rating on Select Method(s)) tends to get especially low ratings, while rubric *InterpRes* (Rating on Interpret Results) and *TxtOrg* (Rating on Text Organization) tends to get especially high ratings. Besides, the distribution of ratings given by each rater is pretty much indistinguishable from the other raters, too, and rater 3 tends to give lower ratings than other 2 raters.

Considering the value of an intraclass correlation that is less than 0.50 as poor reliability, and the value of an intra-class correlation that is between 0.5 and 0.75 as moderate reliability, for the rubric *RsrchQ*, *InitEDA*, *InterpRes* and *TxtOrg*, the raters are inconsistent with one another in how they rate, while, for *CritDes*, *SelMeth* and *VisOrg*, the raters are consistent with one another in how they rate.

From our final model, first, we fitted multi-level models for both the individual 7-rubric models, and then a more "generalized" model. For the individual models, interaction terms and random effects do not appear to be equally important for each rubric, which is reasonable because we are fitting individual models for each of the rubrics. In the more "generalized" model, we can say that ratings are effected by various factors in the experiment, including rater, semester, sex, repeated and rubric. To be more specific, it is mostly influenced by rater, rubric and artifact. On each artifact, the raters are not interpreting the evidence in the same way, and the artifacts are not all of equal quality on each rubric. Moreover, average scores vary from rubric to rubric, and it also varies a bit from one artifact to the next by a small random effect that depends on artifact. Also, each rater also uses each rubric in a way that is not like, or even parallel to, other rater's rubric usage.

Additional analyses and EDA on semester variable were also performed on the data set in order to gain further insight. There is a difference in ratings between the fall and spring semesters, and one possible explanation can be the different tracks that students take the general education courses.

However, there are still some limitations with the final model. The biggest issue was with the random effects and the sample size. First, the data set is small and contains some missing values, which suggests that future studies should sample a larger number of artifacts and a more comprehensive way to deal with these missing values. Moreover, the investigation of the proportion of ratings in the fall and spring semesters shows that there may be other factors that affect ratings. It would be intuitive for additional analyses to consider more variables.

References

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2021). *RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.*

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

Appendix

Naijia Liu

12/8/2021

Contents

Appendix 1. Distribution of Ratings 1
Appendix 2. Agreement Among the Raters 16
Appendix 3. Relationship between Ratings and Various Factors 18
Appendix 4. Anything Else Interesting about Data 33

Appendix 1. Distribution of Ratings

```
library(tidyverse)
library(kableExtra)
library(GGally)
library(grid)
library(gridExtra)
library(ggplotify)
library(reshape2)
library(ggpubr)
library(arm)
library(lme4)
library(caret)
```

```
ratings <- read.csv("/Users/bb/Documents/36-617\ Applied\ Linear\ Model/project02/ratings.csv")
tall <- read.csv("/Users/bb/Documents/36-617\ Applied\ Linear\ Model/project02/tall.csv")
```

```
# take a look at the "head" of all the variables
head(ratings[,1:10]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
head(ratings[,c(1,11:15)]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

We can also check to see how many unique values each variable has.

```
apply(ratings,2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption=" ") %>%
  kable_classic(full_width=F)
```

Table 1:

X	Rater	Sample	Overlap	Semester	Sex	RsrchQ	CritDes	InitEDA	SelMeth
1	3	1	5	Fall	M	3	3	2	2
2	3	2	7	Fall	F	3	3	3	3
3	3	3	9	Spring	F	2	1	3	2
4	3	4	8	Spring	M	2	2	2	1
5	3	5	NA	Fall	–	3	3	3	3
6	3	6	NA	Fall	M	2	1	2	2

Table 2:

X	InterpRes	VisOrg	TxtOrg	Artifact	Repeated
1	2	2	3	O5	1
2	3	3	3	O7	1
3	3	3	3	O9	1
4	1	1	1	O8	1
5	3	3	3	5	0
6	2	2	2	6	0

Table 3:

	unique values
X	117
Rater	3
Sample	117
Overlap	14
Semester	2
Sex	3
RsrchQ	4
CritDes	5
InitEDA	4
SelMeth	3
InterpRes	4
VisOrg	5
TxtOrg	4
Artifact	91
Repeated	2

Indicate missing data (NA's)

We can check for NA's directly:

```
tall[apply(tall,1,function(x){any(is.na(x))}),]
```

```
##       X Rater Artifact Repeated Semester Sex  Rubric Rating
## 161 161     2       45        0      S19   F CritDes     NA
## 684 684     1      100        0      F19   F  VisOrg     NA
```

```
ratings[apply(ratings,1,function(x){any(is.na(x))}),]
```

```
##       X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth
## 5     5     3      5      NA     Fall  --      3       3       3       3
## 6     6     3      6      NA     Fall   M      2       1       2       2
## 7     7     3      7      NA     Fall   F      2       1       3       2
## 8     8     3      8      NA   Spring   F      2       1       2       2
## 9     9     3      9      NA   Spring   F      3       1       2       2
## 13   13     3     13      NA     Fall   F      2       2       1       1
## 14   14     3     15      NA     Fall   M      2       3       3       2
## 15   15     3     16      NA     Fall   F      2       3       4       2
## 16   16     3     17      NA   Spring   F      3       2       2       1
## 20   20     3     21      NA   Spring   M      3       3       4       2
## 21   21     3     22      NA     Fall   F      3       3       3       2
## 22   22     3     23      NA   Spring   F      2       1       1       1
## 23   23     3     24      NA     Fall   F      2       2       2       2
## 24   24     3     25      NA   Spring   F      2       3       2       1
## 25   25     3     26      NA     Fall   M      2       1       2       3
## 26   26     3     27      NA     Fall   M      2       2       3       2
## 27   27     3     28      NA   Spring   M      1       1       1       1
## 31   31     3     32      NA     Fall   M      3       3       3       2
## 32   32     3     33      NA     Fall   M      2       2       3       3
## 33   33     3     34      NA     Fall   M      2       3       2       3
## 34   34     3     35      NA     Fall   F      2       1       2       2
## 35   35     3     36      NA     Fall   F      2       2       2       2
## 36   36     3     37      NA     Fall   M      2       3       2       2
## 37   37     3     38      NA     Fall   M      2       2       2       2
## 38   38     3     39      NA   Spring   F      3       1       3       2
## 39   39     3     40      NA     Fall   M      2       2       3       2
## 44   44     2     45      NA   Spring   F      2      NA       2       2
## 45   45     2     46      NA   Spring   F      2       2       3       2
## 46   46     2     47      NA   Spring   M      3       3       2       1
## 47   47     2     48      NA     Fall   M      3       3       4       3
## 48   48     2     49      NA     Fall   M      3       1       3       2
## 52   52     2     53      NA     Fall   F      3       4       3       3
## 53   53     2     54      NA     Fall   M      1       1       3       2
## 54   54     2     55      NA     Fall   F      3       2       2       2
## 55   55     2     56      NA     Fall   F      2       3       2       2
## 56   56     2     57      NA     Fall   M      2       1       2       2
## 60   60     2     61      NA     Fall   M      2       2       3       2
## 61   61     2     62      NA   Spring   F      3       4       4       2
## 62   62     2     63      NA   Spring   F      3       3       3       2
## 63   63     2     64      NA     Fall   F      2       3       2       2
```

```
## 64   64     2     65      NA     Fall   F      3       3       2       2
## 65   65     2     66      NA   Spring   F      3       3       4       2
## 66   66     2     67      NA     Fall   F      3       1       3       2
## 67   67     2     68      NA   Spring   M      3       3       2       2
## 71   71     2     72      NA   Spring   F      2       2       3       2
## 72   72     2     73      NA     Fall   F      2       1       1       2
## 73   73     2     74      NA     Fall   M      2       1       3       3
## 74   74     2     75      NA     Fall   F      2       2       3       2
## 75   75     2     76      NA     Fall   M      2       2       2       2
## 76   76     2     77      NA     Fall   M      2       2       2       2
## 77   77     2     78      NA     Fall   M      3       3       3       3
## 78   78     2     79      NA     Fall   M      3       2       3       3
## 83   83     1     84      NA   Spring   M      3       2       2       2
## 84   84     1     85      NA     Fall   M      4       3       3       2
## 85   85     1     86      NA   Spring   F      3       2       2       2
## 86   86     1     87      NA     Fall   M      3       2       1       2
## 87   87     1     88      NA   Spring   F      3       3       3       2
## 91   91     1     92      NA     Fall   F      3       1       2       2
## 92   92     1     93      NA   Spring   F      3       1       2       2
## 93   93     1     94      NA     Fall   F      3       3       4       2
## 94   94     1     95      NA     Fall   M      2       2       3       3
## 95   95     1     96      NA     Fall   F      3       2       3       2
## 99   99     1    100      NA     Fall   F      2       3       2       3
## 100 100     1    101      NA   Spring   F      1       1       3       2
## 101 101     1    102      NA     Fall   M      1       1       2       2
## 102 102     1    103      NA     Fall   M      2       2       3       3
## 103 103     1    104      NA     Fall   F      2       1       3       2
## 104 104     1    105      NA     Fall   M      2       1       2       2
## 105 105     1    106      NA     Fall   M      3       2       1       2
## 106 106     1    107      NA     Fall   M      3       1       2       2
## 110 110     1    111      NA   Spring   F      2       1       2       2
## 111 111     1    112      NA     Fall   M      3       1       3       2
## 112 112     1    113      NA   Spring   F      2       1       1       2
## 113 113     1    114      NA   Spring   F      2       1       3       2
## 114 114     1    115      NA   Spring   F      3       1       3       2
## 115 115     1    116      NA     Fall   F      2       1       2       2
## 116 116     1    117      NA     Fall   F      2       1       2       2
## 117 117     1    118      NA     Fall   F      2       1       2       2
##     InterpRes VisOrg TxtOrg Artifact Repeated
## 5           3      3      3        5        0
## 6           2      2      2        6        0
## 7           2      2      2        7        0
## 8           2      2      2        8        0
## 9           2      2      2        9        0
## 13          1      1      1       13        0
## 14          2      4      3       15        0
## 15          3      3      4       16        0
## 16          2      2      2       17        0
## 20          3      3      4       21        0
## 21          3      2      3       22        0
## 22          1      1      1       23        0
## 23          2      2      3       24        0
## 24          1      2      2       25        0
## 25          2      2      2       26        0
```

```
## 26          2      2      3       27         0
## 27          1      1      1       28         0
## 31          3      3      3       32         0
## 32          2      2      2       33         0
## 33          3      2      2       34         0
## 34          2      3      2       35         0
## 35          2      2      3       36         0
## 36          2      2      3       37         0
## 37          2      3      2       38         0
## 38          2      2      2       39         0
## 39          2      2      2       40         0
## 44          2      2      3       45         0
## 45          2      2      2       46         0
## 46          2      1      1       47         0
## 47          3      2      4       48         0
## 48          3      4      3       49         0
## 52          3      3      3       53         0
## 53          2      2      3       54         0
## 54          3      3      2       55         0
## 55          2      3      2       56         0
## 56          2      3      3       57         0
## 60          2      3      4       61         0
## 61          3      4      3       62         0
## 62          3      3      3       63         0
## 63          3      3      3       64         0
## 64          1      3      3       65         0
## 65          3      3      2       66         0
## 66          3      3      1       67         0
## 67          3      3      3       68         0
## 71          3      3      2       72         0
## 72          2      2      3       73         0
## 73          3      3      2       74         0
## 74          3      2      2       75         0
## 75          2      2      2       76         0
## 76          3      3      2       77         0
## 77          3      3      3       78         0
## 78          3      3      3       79         0
## 83          3      3      3       84         0
## 84          3      3      3       85         0
## 85          3      2      3       86         0
## 86          2      2      3       87         0
## 87          3      4      3       88         0
## 91          3      2      3       92         0
## 92          3      2      4       93         0
## 93          3      3      3       94         0
## 94          3      2      3       95         0
## 95          3      3      3       96         0
## 99          3     NA      2      100         0
## 100         2      3      3      101         0
## 101         3      2      1      102         0
## 102         3      2      3      103         0
## 103         3      2      3      104         0
## 104         2      2      2      105         0
## 105         2      2      3      106         0
```

Table 4:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
RsrchQ	1	2	2	2.36	3.0	4	0.60
CritDes	1	1	2	1.86	2.5	4	0.84
InitEDA	1	2	2	2.44	3.0	4	0.70
SelMeth	1	2	2	2.06	2.0	3	0.48
InterpRes	1	2	3	2.49	3.0	4	0.61
VisOrg	1	2	2	2.42	3.0	4	0.68
TxtOrg	1	2	3	2.60	3.0	4	0.70

```
## 106          3       2       2       107          0
## 110          2       2       2       111          0
## 111          3       3       3       112          0
## 112          3       2       2       113          0
## 113          3       2       3       114          0
## 114          3       3       3       115          0
## 115          2       3       3       116          0
## 116          3       4       3       117          0
## 117          3       3       3       118          0
```

There appears to be missing values in "Overlap", the rubric `CritDes` and `VisOrg`. For models involving five of the rubrics we will get all the data from all the raters, but for models involving `CritDes` we would be missing a rating from Rater 2, and for models involving `VisOrg` we would be missing a rating from Rater 1. Since they could undermine some model comparisons, we will delete data from row 44 and row 99 for numeric summary.

Also, note that none of the missing values occur in the smaller 13-rubric data set, since none of the artifact that has missing value is repeated. So we don't have to worry about missing data at all in analyses that just involve this smaller data set.

Moreover, we will also have to be careful of the missing "Sex" value, which is currently coded as "–".

```
# convert "NA" Overlap into 0
ratings$Overlap[which(is.na(ratings$Overlap))] <- 0
```

distribution of ratings for each rubrics

Next, let's make a table with the usual one-dimensional summary statistics for each rubric.

```
ratings_rubric <- ratings[-c(44,99),c(7:13)] ## extract data only for 7 rubrics

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

For the table above, we might also check for old-fashioned missing value codes like "9", "99", "98", etc., but there's no evidence of that. (look at the Min and Max values - no "9's", "99's", etc.))

```
ggplot(gather(ratings_rubric), aes(value)) +
  geom_histogram(bins=10) +
  facet_wrap(~key, scales = 'free_x')
```

```
# table of counts for each rubric across 3 raters
tall$Rating <- factor(tall$Rating,levels=1:4)

for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}

tmp0 <- lapply(split(tall$Rating,tall$Rubric),summary)
tmp <- data.frame(matrix(0,nrow=5,ncol=7))  ## seven rubrics...
names(tmp) <- names(tmp0)
row.names(tmp) <- c(paste("Rating",1:4),"<NA>")
for (i in names(tmp0)) {
  tmp[,i] <- tmp[,i] + c(tmp0[[i]],0)[1:5]
}

tmp
```

```
##          CritDes InitEDA InterpRes RsrchQ SelMeth TxtOrg VisOrg
## Rating 1      47       8         6      6      10      8      7
## Rating 2      39      56        49     65      89     37     59
## Rating 3      28      47        61     45      18     66     45
## Rating 4       2       6         1      1       0      6      5
## <NA>           1       0         0      0       0      0      1
```

We assumed that ratings with values that are less than or equal to 2 for each rubrics are considered as low, and that ratings with values that are higher than 2 for each rubrics are considered as high.

1) Rubric "SelMeth" (Rating on Select Method(s)) tend to get especially low ratings.

That is because the 3rd quantile for "SelMeth" is 2, which is the lowest among all the rubrics. This means that at least 75% of artifacts get score lower than 2 for rubric "SelMeth". And the max score for rubric "Selmeth" is 3, which is also lower than all the other rubrics. We can also see from the table that 99 artifacts get score that is equal or lower than 2, which collides with our findings in the histogram.

Though, from the histogram, the percentage that artifacts scored in 1 of the rubric "CritDes" is the lowest among all the other rubrics and has largest number of rating 1, the total amount of artifacts scored less than or equal to 2 is lower than "SelMeth". So, we are still apt to draw the conclusion that rubric "SelMeth" tend to get especially low ratings.

2) Rubric "InterpRes" (Rating on Interpret Results) and "TxtOrg" (Rating on Text Organization) tend to get especially high ratings.

That is because the median for both rubrics is 3, which is higher than the others. This implies that at least 50% of artifacts get score higher than 3 for these two rubrics. The mean for "InterpRes" is 2.49 and the mean for "TxtOrg" is 2.60, which are the top two highest among all rubrics. We can also see from the histogram that over 60 of artifacts score higher or equal to 3, which correspnds to what we have found in the statistics summary table.

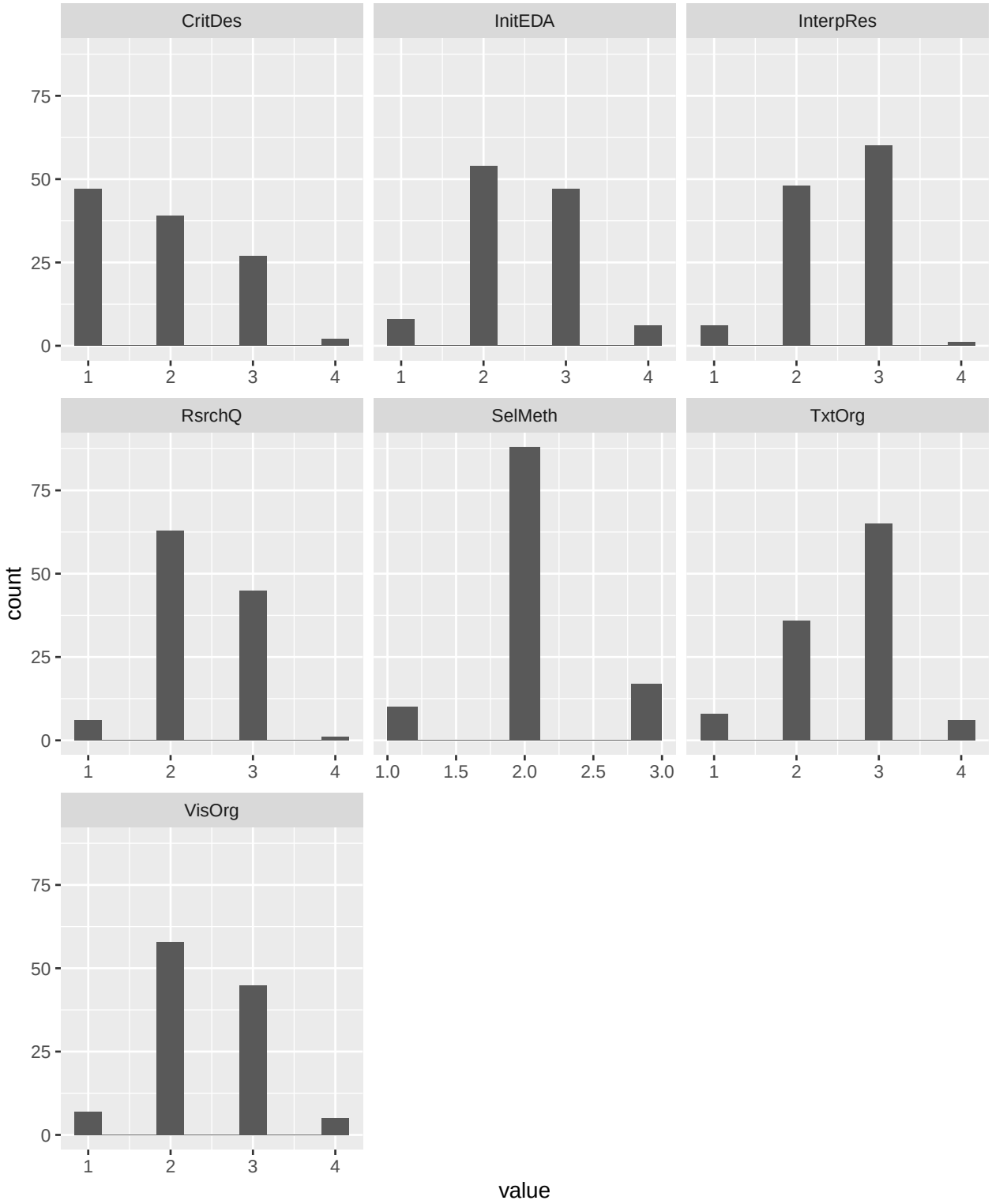


Figure 1: Distributions of Rubrics

Table 5:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
Rater1	1	2	2	2.35	3	4	0.70
Rater2	1	2	2	2.43	3	4	0.70
Rater3	1	2	2	2.18	3	4	0.69

distribution of ratings given by each rater

Now, those NA's have me curious...

```
tall[apply(tall,1,function(x){any(is.na(x))}),]
```

```
##       X Rater Artifact Repeated Semester Sex  Rubric Rating
## 161 161     2       45        0      S19   F CritDes   <NA>
## 684 684     1      100        0      F19   F  VisOrg   <NA>
```

```
ratings[ratings$Sex=="--",]
```

```
##   X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth InterpRes
## 5 5     3      5       0     Fall  --      3       3       3       3         3
##   VisOrg TxtOrg Artifact Repeated
## 5      3      3        5        0
```

Same as whta we have done before, the value missing for `Rating`' will be deleted from row 161 and row 684 for numeric summary. Next, let's make a table with the usual one-dimensional summary statistics for each rubric.

```
tall <- read.csv("/Users/bb/Documents/36-617\ Applied\ Linear\ Model/project02/tall.csv")

ratings_rater <- tall[-c(161,684),c(2,8)]
## extract data only for rubrics without missing values

# make 3 rating subsets for each rater
ratings_rater1 <- ratings_rater[which(ratings_rater$Rater==1),]
ratings_rater2 <- ratings_rater[which(ratings_rater$Rater==2),]
ratings_rater3 <- ratings_rater[which(ratings_rater$Rater==3),]

# statistics summary for all raters
r <- cbind(c(summary(ratings_rater1[,2]),SD=sd(ratings_rater1[,2])),
           c(summary(ratings_rater2[,2]),SD=sd(ratings_rater2[,2])),
           c(summary(ratings_rater3[,2]),SD=sd(ratings_rater3[,2]))) %>%
  as.data.frame

colnames(r) <- c("Rater1","Rater2","Rater3")

r %>% t() %>% round(digits=2) %>% kbl(booktabs=T,caption=" ") %>%
  kable_classic()
```

```
rater1 <- ggplot(data=ratings_rater1,aes(Rating)) +
  geom_histogram(bins=10) + ylim(c(0,150))

rater2 <- ggplot(data=ratings_rater2,aes(Rating)) +
  geom_histogram(bins=10) + ylim(c(0,150))

rater3 <- ggplot(data=ratings_rater3,aes(Rating)) +
  geom_histogram(bins=10) + ylim(c(0,150))


ggarrange(rater1, rater2, rater3,
          labels = c("Rater1", "Rater2", "Rater3"),
          ncol = 3, nrow = 1)
```

```
# the table of counts across raters
tall$Rating <- factor(tall$Rating,levels=1:4)

for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}

tmp0 <- lapply(split(tall$Rating,tall$Rater),summary)
tmp <- data.frame(matrix(0,nrow=5,ncol=3))  ## three raters...
names(tmp) <- names(tmp0)
row.names(tmp) <- c(paste("Rating",1:4),"<NA>")

for (i in names(tmp0)) {
  tmp[,i] <- tmp[,i] + c(tmp0[[i]],0)[1:5]
}
names(tmp) <- paste("Rater",1:3)

tmp
```

```
##          Rater 1 Rater 2 Rater 3
## Rating 1      29      23      40
## Rating 2     125     119     150
## Rating 3     112     120      78
## Rating 4       6      10       5
## <NA>           1       1       0
```

We assumed that ratings with values that are less than or equal to 2 for each rubrics are considered as low, and that ratings with values that are higher than 2 for each rubrics are considered as high.

1) From the statistics summary table above, the mean for all 3 raters are pretty similar, which are 2.35, 2.43 and 2.18, and the SD for all 3 raters are pretty similar too, which are all around 0.7. Thus, the distribution of ratings given by each rater is pretty much indistinguishable from the other raters.

2) However, from the histogram and the corresponding table of counts above, it seems that rater3 tend to give lower ratings than other 2 raters.

That is because rater 3 have the highest percent of scoring 2, and it has the highest total number of ratings that are equal or lower than 2 among all 3 raters. Though, the distribution of ratings given by rater 1 and rater 2 is pretty much indistinguishable from each other.

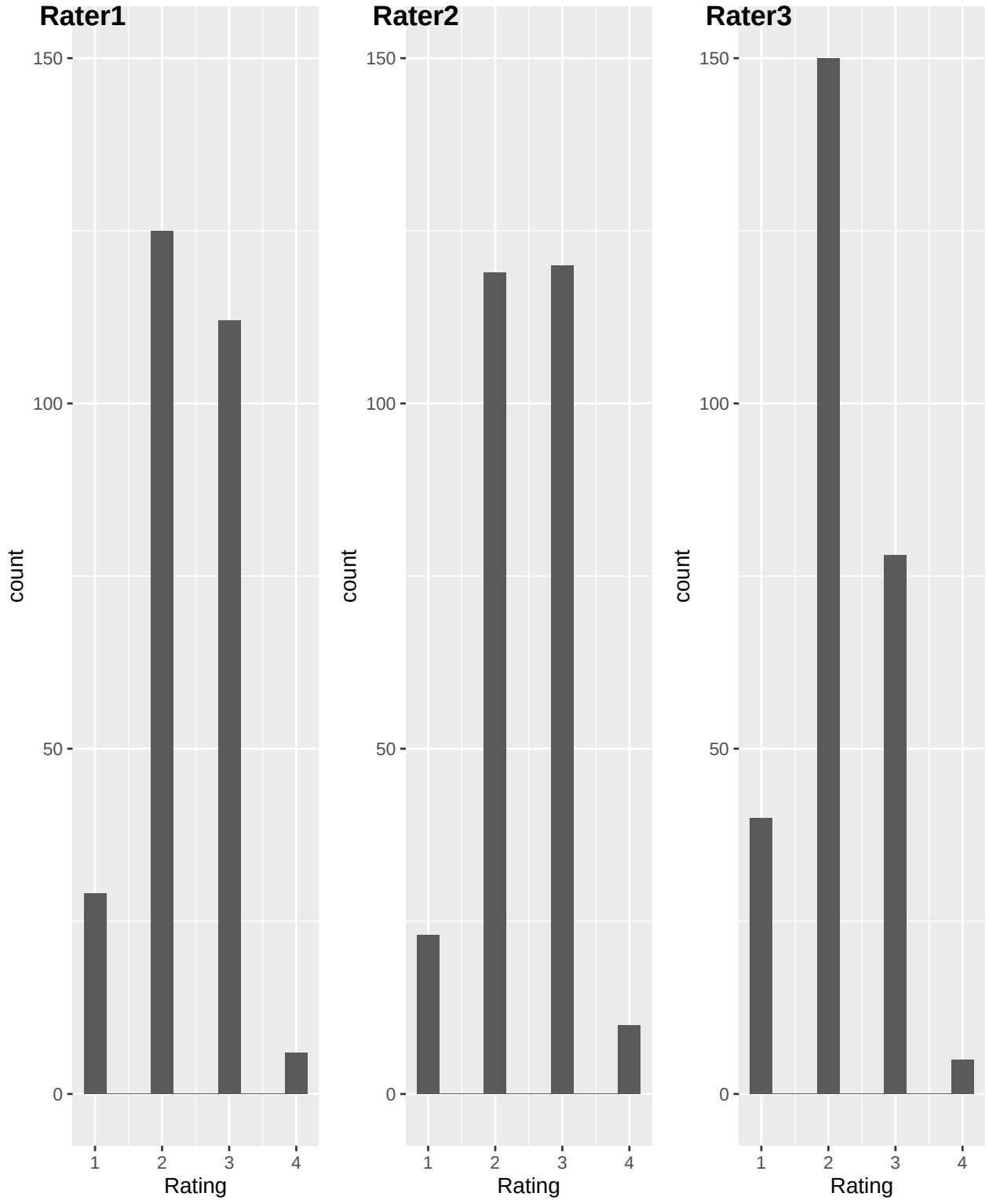


Figure 2: Distributions of Rubrics by Raters

Table 6:

	X	Rater	Sample	Overlap	Semester	Sex	RsrchQ	CritDes	InitEDA	SelMeth
1	1	3	1	5	Fall	M	3	3	2	2
2	2	3	2	7	Fall	F	3	3	3	3
3	3	3	3	9	Spring	F	2	1	3	2
4	4	3	4	8	Spring	M	2	2	2	1
10	10	3	10	10	Fall	F	2	1	2	2
11	11	3	11	13	Fall	M	2	2	2	2

Table 7:

	X	InterpRes	VisOrg	TxtOrg	Artifact	Repeated
1	1	2	2	3	O5	1
2	2	3	3	3	O7	1
3	3	3	3	3	O9	1
4	4	1	1	1	O8	1
10	10	3	2	3	O10	1
11	11	2	3	3	O13	1

distribution of the rubrics for the 13 artifacts subset

Now, we will see whether 13 artifacts are representative of the whole set of 91 artifacts.

```
ratings <- read.csv("/Users/bb/Documents/36-617\ Applied\ Linear\ Model/project02/ratings.csv")

# make a subset of the data for just the 13 artifacts
ratings13 <- ratings[which(ratings$Repeated==1),]
```

```
# take a look at the "head" of all the variables
head(ratings13[,1:10]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
head(ratings13[,c(1,11:15)]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

As has mentioned before, there is no there is no NA's for the subset of 13 artifacts. Next, let's make a table with the usual one-dimensional summary statistics for each rubric.

```
ratings13_rubric <- ratings13[,c(7:13)] ## extract data only for 7 rubrics

apply(ratings13_rubric,2,function(x) c(summary(x),SD=sd(x))) %>%
  as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

For the table above, we might also check for old-fashioned missing value codes like "9", "99", "98", etc., but there's no evidence of that. (look at the Min and Max values - no "9's", "99's", etc.))

Table 8:

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

```
ggplot(gather(ratings13_rubric), aes(value)) +
  geom_histogram(bins=10) +
  facet_wrap(~key, scales = 'free_x')
```

```
tall.13 <-  tall[grep("O",tall$Artifact),]

#  make the title of each facet
rater.name <- function(x) { paste("Rater",x) }

## Barplots for reduced data...
g <- ggplot(tall.13,aes(x = Rating)) +
  facet_wrap( ~ Rater, labeller=labeller(Rater=rater.name)) +
  geom_bar()

g
```

```
tall$Rating <- factor(tall$Rating,levels=1:4)
for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}

ratings.13 <- ratings[grep("O",ratings$Artifact),]
tall.13 <-  tall[grep("O",tall$Artifact),]

# Table of counts
tmp <- data.frame(lapply(split(tall.13$Rating,tall.13$Rubric),summary))
row.names(tmp) <- paste("Rating",1:4)

tmp
```

```
##          CritDes InitEDA InterpRes RsrchQ SelMeth TxtOrg VisOrg
## Rating 1      17       1         1      2       4      2      3
## Rating 2      16      22        18     24      29     10     22
## Rating 3       6      16        19     13       6     26     14
## Rating 4       0       0         1      0       0      1      0
```

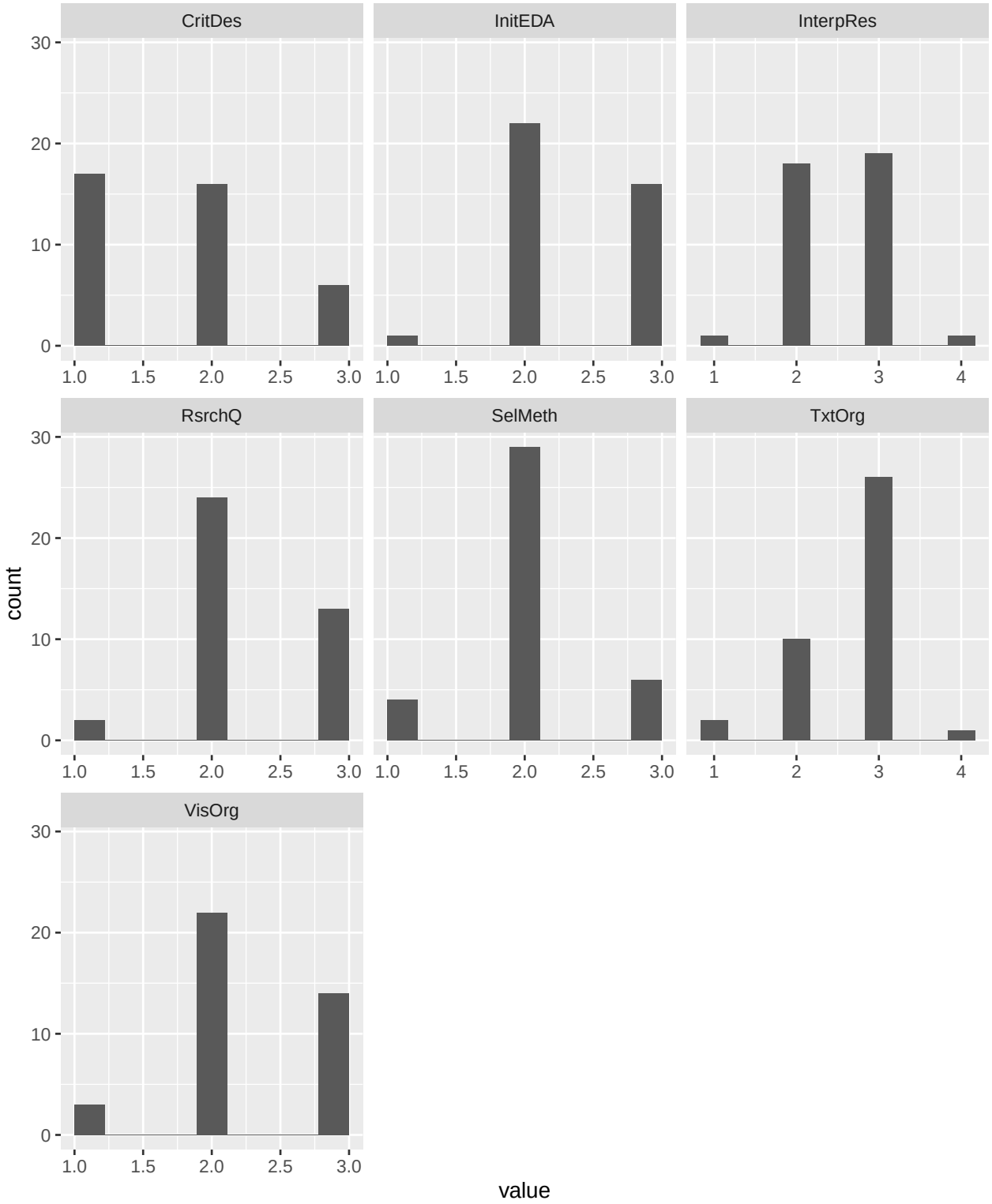


Figure 3: Distributions of Subset Rubrics

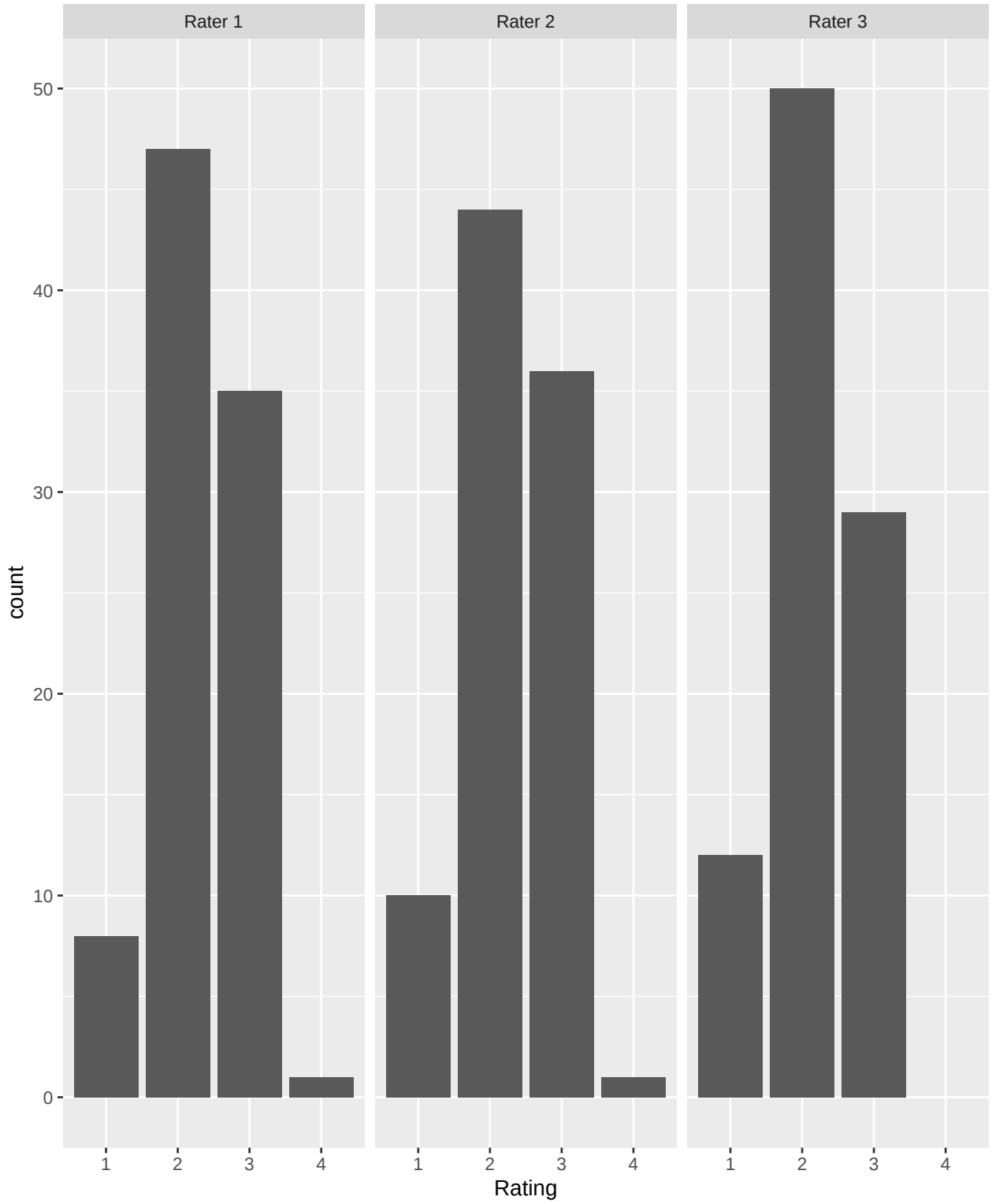


Figure 4: Distributions of Subset Raters

```
# Corresponding table of counts...
tmp <- data.frame(lapply(split(tall.13$Rating,tall.13$Rater),summary))
row.names(tmp) <- paste("Rating",1:4)
names(tmp) <- paste("Rater",1:3)

tmp
```

```
##          Rater 1 Rater 2 Rater 3
## Rating 1       8      10      12
## Rating 2      47      44      50
## Rating 3      35      36      29
## Rating 4       1       1       0
```

Yes, these 13 artifacts are representative of the whole set of 91 artifacts, because:

1) From the statistics summary above, the 3rd quantile, the mean and the standard deviation of all 7 rubrics of the whole set of 91 artifacts are pretty similar to those of the 13 artifacts subsets. Minimum value, 1st quantile and the median are same in both data set. However, the max values of all 7 rubrics of the whole set of 91 artifacts are different from those of the 13 artifacts subsets.

2) From the histogram and corresponding counts table above, the distribution of ratings for each rubrics in the 13 artifacts is pretty much indistinguishable from the whole set of 91 artifacts. Though, there are only 3 bars left in 6 of 7 rubrics of the 13 artifacts subset.

Appendix 2. Agreement Among the Raters

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

```
# useful preliminaries
Rubric.names <- sort(unique(tall$Rubric))

ICC.vec <- NULL
for (i in Rubric.names) {

  tmp <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=tall.13[tall.13$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.vec <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names

agreement.results <- cbind(ICC.common=ICC.vec,"       a12"=0,a23=0,a13=0)

agreement.tables <- as.list(rep(NA,7))
names(agreement.tables) <- Rubric.names

for (i in Rubric.names) {
  r12 <- data.frame(r1=factor(ratings.13[ratings.13$Rater==1,i],levels=1:4),
                    r2=factor(ratings.13[ratings.13$Rater==2,i],levels=1:4),
```

```
                    a1=ratings.13[ratings.13$Rater==1,"Artifact"],
                    a2=ratings.13[ratings.13$Rater==2,"Artifact"])
  if(any(r12[,3]!=r12[,4])) { stop(paste("Rater 1-2 Artifact mismatch on rubric",i)) }
  a12 <- mean(r12[,1]==r12[,2])
  r12 <- table(r12[,1:2])  ## print this to see how much agreement there is among raters 1-2

  r23 <- data.frame(r2=factor(ratings.13[ratings.13$Rater==2,i],levels=1:4),
                    r3=factor(ratings.13[ratings.13$Rater==3,i],levels=1:4),
                    a2=ratings.13[ratings.13$Rater==2,"Artifact"],
                    a3=ratings.13[ratings.13$Rater==3,"Artifact"])
  if(any(r23[,3]!=r23[,4])) { stop(paste("Rater 2-3 Artifact mismatch on rubric",i)) }
  a23 <- mean(r23[,1]==r23[,2])
  r23 <- table(r23[,1:2])  ## print this to see how much agreement there is among raters 2-3

  r13 <- data.frame(r1=factor(ratings.13[ratings.13$Rater==1,i],levels=1:4),
                    r3=factor(ratings.13[ratings.13$Rater==3,i],levels=1:4),
                    a1=ratings.13[ratings.13$Rater==1,"Artifact"],
                    a3=ratings.13[ratings.13$Rater==3,"Artifact"])
  if(any(r13[,3]!=r13[,4])) { stop(paste("Rater 1-3 Artifact mismatch on rubric",i)) }
  a13 <- mean(r13[,1]==r13[,2])
  r13 <- table(r13[,1:2])  ## print this to see how much agreement there is among raters 1-3

  agreement.results[i,2:4] <- c(a12,a23,a13)

  agreement.tables[[i]] <- list(r12,r23,r13)

}


ICC.vec <- NULL
for (i in Rubric.names) {

  tmp <- lmer(as.numeric(Rating) ~ 1 + (1|Artifact), data=tall[tall$Rubric==i,])
  sig2 <- summary(tmp)$sigma^2
  tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
  ICC <- tau2 / (tau2 + sig2)
  ICC.vec <- c(ICC.vec,ICC)
}
names(ICC.vec) <- Rubric.names

agreement.results <- cbind(ICC.alldata=ICC.vec,agreement.results)

round(agreement.results,2)
```

```
##           ICC.alldata ICC.common        a12  a23  a13
## CritDes          0.67       0.57       0.54 0.69 0.62
## InitEDA          0.69       0.49       0.69 0.85 0.54
## InterpRes        0.22       0.23       0.62 0.62 0.54
## RsrchQ           0.21       0.19       0.38 0.54 0.77
## SelMeth          0.47       0.52       0.92 0.69 0.62
## TxtOrg           0.19       0.14       0.69 0.54 0.62
## VisOrg           0.66       0.59       0.54 0.77 0.77
```

First we examine the 13 "common" artifacts that all 3 raters saw. As we consider the value of an intra-class correlation that is less than 0.50 as poor reliability, and the value of an intra-class correlation that is between 0.5 and 0.75 as moderate reliability.

Thus, we could say that, based on the rules of thumb for interpreting ICC, the rubric `RsrchQ`, `InitEDA`, `InterpRes` and `TxtOrg` can be rated with "poor" reliability by different raters, which means the raters are inconsistent with one another in how they rate; while, `CritDes`, `SelMeth` and `VisOrg` can be can be rated with "good" reliability by different raters, which means the raters are consistent with one another in how they rate. The higher ICC of the rubric is, the more raters agree.

So, now we will make a 2-way table of counts for the ratings of each pair of raters on each rubric to tell which raters might be contributing to disagreement.

The percentage of observations for rubric `InitEDA` with Rater1 and Rater3 is the lowest among all 3 pairs, which implies the lowest agreement between the two raters. This means on rubric `InitEDA`, Rater1 and Rater3 are disagreeing most with each other.

The percentage of observations for rubric `InterpRes` with Rater1 and Rater3 is the lowest among all 3 pairs, which implies the lowest agreement between the two raters. This means on rubric `InterpRes`, Rater1 and Rater3 are disagreeing most with each other.

The percentage of observations for rubric `RsrchQ` with Rater1 and Rater2 is the lowest among all 3 pairs, which implies the lowest agreement between the two raters. This means on rubric `RsrchQ`, Rater1 and Rater2 are disagreeing most with each other most.

The percentage of observations for rubric `TxtOrg` with Rater2 and Rater3 is the lowest among all 3 pairs, which implies the lowest agreement between the two raters. This means on rubric `TxtOrg`, Rater2 disagree most with both Rater3.

For the ICC calculations on the full data set, we do not have the percent exact agreement calculations on the full data set, that is because the other 78 artifacts are only graded by only one grader, and there is no way to compare the rating between any 2 raters on the same rubric.

The seven ICC's for the full data set agree with the seven ICC's for the subset corresponding to the 13 artifacts that all three raters. As we can see, the difference between the seven ICC's for the full data set and the seven ICC's for the subset is relatively small.

Appendix 3. Relationship between Ratings and Various Factors

3(i): Adding fixed effects to the seven rubric-specific models using just the data from the 13 common artifacts that all three raters saw

```
library(RLRsim)

library(LMERConvenienceFunctions, warn.conflicts=F, quietly=T)
library(lme4, warn.conflicts=F, quietly=T)

Rubric.names <- sort(unique(tall$Rubric))

model.formula.13 <- as.list(rep(NA,7))
names(model.formula.13) <- Rubric.names

for (i in Rubric.names) {

  ## fit each base model
  rubric.data <- tall.13[tall.13$Rubric==i,]
```

```
  tmp <- 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
  }

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

}
```

The final model we got for each rubric based on the 13 common artifacts that all three raters saw.

```
## see what "final models" we go for each rubric
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)
```

So, it looks like we don't need to add any fixed effects or interactions to the models for each rubric, using only the data reduced to the 13 common rubrics.

3(ii): Adding fixed effects to the seven rubric-specific models using all the data

Now let's try with the full data.

```
tall <- read.csv("/Users/bb/Documents/36-617\ Applied\ Linear\ Model/project02/tall.csv")

tall$Rating <- factor(tall$Rating,levels=1:4)
for (i in unique(tall$Rubric)) {
  ratings[,i] <- factor(ratings[,i],levels=1:4)
}

tall$Sex[nchar(tall$Sex)==0] <- "--"


Rubric.names <- sort(unique(tall$Rubric))

# delete the rows with missing ratings
tall.nonmissing <- tall[-c(161,684),]

#since there is no good justification for how to impute the "Sex" of the student
# eliminate that person from the data set

tall.nonmissing <- tall.nonmissing[tall.nonmissing$Sex!="--",]

model.formula.alldata <- as.list(rep(NA,7))
names(model.formula.alldata) <- Rubric.names


for (i in Rubric.names) {

  ## fit each base model
  rubric.data <- tall.nonmissing[tall.nonmissing$Rubric==i,]
  tmp <- 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
  }

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

}
```

The final model we got for each rubric based on the full data set.

```
## see what "final models" we got...
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) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
##
## $TxtOrg
## as.numeric(Rating) ~ (1 | Artifact)
##
## $VisOrg
## as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

3(iii): Trying interactions and new random effects for the seven rubric specific models using all the data

Now we see there are some differences among the models: For `InitEDA`, `RsrchQ` and `TxtOrg`, the models are just the simple random-intercept models. For the other four, the models are a little more complex. We should examine each of these 4 models to see (a) if the fixed effects make sense to us; and (b) if there are any interactions or additional random effects to consider.

First, for rubric `SelMeth`.

```
# refit the model and check on the t-statistics
fla <- formula(model.formula.alldata[["SelMeth"]])
tmp <- lmer(fla,data=tall.nonmissing[tall.nonmissing$Rubric=="SelMeth",])
round(summary(tmp)$coef,2)
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     2.25       0.08   29.99
## as.factor(Rater)2     2.23       0.07   29.99
## as.factor(Rater)3     2.03       0.08   27.03
## SemesterS19          -0.36       0.10   -3.66
```

```
# now check to make sure we really need "Rater" as a factor...
tmp.single_intercept <- update(tmp, . ~ . + 1 - as.factor(Rater))
anova(tmp.single_intercept,tmp)
```

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

```
## Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
## Models:
## tmp.single_intercept: as.numeric(Rating) ~ Semester + (1 | Artifact)
## tmp: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp.single_intercept    4 145.07 156.08 -68.534   137.07
## tmp                     6 142.05 158.58 -65.027   130.05 7.0146  2    0.02998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# model "tmp" is preferred

# add fixed-effect interactions
tmp.fixed_interactions <- update(tmp, . ~ . + as.factor(Rater)*Semester - Semester)
anova(tmp,tmp.fixed_interactions)
```

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

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

```
# model "tmp" is preferred
```

```
# check for random effects.
# Testing (Semester|Artifact)...

#m0 <- tmp                                          ## Null hypothesis
#mA <- update(m0, . ~ . + (Semester|Artifact))      ## Alternative hypotheses
#m  <- update(mA, . ~ . - (1|Artifact))             ## Model with only the new R.E.

#exactRLRT(m0=m0,mA=mA,m=m)

# for model mA is: there are more random effects than there are observations
# in the data set. Thus, the model isn't even possible, so no testing is needed.

# Testing (as.factor(Rater)|Artifact)

#m0 <- tmp                                          ## Null hypothesis
#mA <- update(m0, . ~ . + (as.factor(Rater)|Artifact))   ## Alternative hypotheses
#m  <- update(mA, . ~ . - (1|Artifact))             ## Model with only the new R.E.

#exactRLRT(m0=m0,mA=mA,m=m)
# for model mA is: there are more random effects than there are observations
# in the data set. Thus, the model isn't even possible, so no testing is needed.
```

```
# so this is our final model for SelMeth:
summary(tmp)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + Semester + (1 | Artifact) -
##     1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "SelMeth", ]
##
## REML criterion at convergence: 143.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0480 -0.3923 -0.0551  0.2674  2.5827
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08973  0.2996
##  Residual             0.10842  0.3293
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.25037    0.07503  29.992
## as.factor(Rater)2  2.22653    0.07424  29.991
## as.factor(Rater)3  2.03316    0.07521  27.033
## SemesterS19       -0.35860    0.09796  -3.661
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2 a.(R)3
## as.fctr(R)2  0.285
## as.fctr(R)3  0.287  0.280
## SemesterS19 -0.413 -0.391 -0.394
```

Considering the same rater in the same semester (ex. Semester Spring 19), the rating for rubric `SelMeth` is distinguishable in different artifacts. For the same artifact in the same semester, rater 1 and rater 2 tend to give the similar rating for rubric `SelMeth`, while rater 3 gives the rating that is 0.19337 lower than rater 2, and 0.21721 than rater 1. For the same artifact rated by the same rater, the rating for rubric `SelMeth` in Semester Spring 19 is 0.35860 lower than that in Semester Fall 19.

Next, for rubric `CritDes`,`InterpRes` and `VisOrg`, since there is just one fixed-effect, we will only try to add random effects.

```
## refit the model and check on the t-statistics
fla1 <- formula(model.formula.alldata[["CritDes"]])
tmp1 <- lmer(fla1,data=tall.nonmissing[tall.nonmissing$Rubric=="CritDes",])
round(summary(tmp1)$coef,2)
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     1.69       0.12   13.98
## as.factor(Rater)2     2.11       0.12   17.34
## as.factor(Rater)3     1.89       0.12   15.51
```

```
fla2 <- formula(model.formula.alldata[["InterpRes"]])
tmp2 <- lmer(fla2,data=tall.nonmissing[tall.nonmissing$Rubric=="InterpRes",])
round(summary(tmp1)$coef,2)
```

```
##                   Estimate Std. Error t value
```

```
## as.factor(Rater)1     1.69       0.12   13.98
## as.factor(Rater)2     2.11       0.12   17.34
## as.factor(Rater)3     1.89       0.12   15.51
```

```
fla3 <- formula(model.formula.alldata[["VisOrg"]])
tmp3 <- lmer(fla3,data=tall.nonmissing[tall.nonmissing$Rubric=="VisOrg",])
round(summary(tmp3)$coef,2)
```

```
##                   Estimate Std. Error t value
## as.factor(Rater)1     2.38        0.1   24.62
## as.factor(Rater)2     2.65        0.1   27.70
## as.factor(Rater)3     2.28        0.1   23.64
```

```
## now check to make sure we really need "Rater" as a factor...
tmp1.single_intercept <- update(tmp1, . ~ . + 1 - as.factor(Rater))
anova(tmp1.single_intercept,tmp1)
```

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
## Models:
## tmp1.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp1: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp1.single_intercept    3 277.68 285.91 -135.84   271.68
## tmp1                     5 273.62 287.35 -131.81   263.62 8.0535  2    0.01783
##
## tmp1.single_intercept
## tmp1                  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## for rubric "InterpRes", model with 'Rater' is preferred

tmp2.single_intercept <- update(tmp2, . ~ . + 1 - as.factor(Rater))
anova(tmp2.single_intercept,tmp2)
```

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
## Models:
## tmp2.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp2: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                       npar    AIC    BIC   logLik deviance  Chisq Df Pr(>Chisq)
## tmp2.single_intercept    3 218.53 226.79 -106.263   212.53
## tmp2                     5 200.66 214.43  -95.331   190.66 21.864  2  1.787e-05
##
## tmp2.single_intercept
## tmp2                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## for rubric "CritDes", model with 'Rater' is preferred

tmp3.single_intercept <- update(tmp3, . ~ . + 1 - as.factor(Rater))
anova(tmp3.single_intercept,tmp3)
```

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

## Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
## Models:
## tmp3.single_intercept: as.numeric(Rating) ~ (1 | Artifact)
## tmp3: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## tmp3.single_intercept    3 227.21 235.44 -110.60   221.21
## tmp3                     5 220.82 234.54 -105.41   210.82 10.392  2   0.005539
##
## tmp3.single_intercept
## tmp3                  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## for rubric "VisOrg", model with 'Rater' is preferred
```

```
## Finally, check for random effects.
## Testng (as.factor(Rater)|Artifact)
## for rubric "InterpRes"
#m10 <- tmp1                                          ## Null hypothesis
#m1A <- update(m10, . ~ . + (as.factor(Rater)|Artifact))   ## Alternative hypotheses
#m1  <- update(m1A, . ~ . - (1|Artifact))          ## Model with only the new R.E.

#exactRLRT(m10=m10,m1A=m1A,m1=m1)

## for rubric "CritDes"
#m20 <- tmp2                                          ## Null hypothesis
#m2A <- update(m20, . ~ . + (as.factor(Rater)|Artifact))   ## Alternative hypotheses
#m2  <- update(m2A, . ~ . - (1|Artifact))          ## Model with only the new R.E.

#exactRLRT(m20=m20,m2A=m2A,m2=m2)

## for rubric "VisOrg"
#m30 <- tmp3                                          ## Null hypothesis
#m3A <- update(m30, . ~ . + (as.factor(Rater)|Artifact))   ## Alternative hypotheses
#m3  <- update(m3A, . ~ . - (1|Artifact))          ## Model with only the new R.E.

#exactRLRT(m30=m30,m3A=m3A,m3=m3)

## for all 3 rubrics, model with random effects isn't even possible,
## since there are more random effects than observations in the data set
```

```
## so this are our final model for "InterpRes", "CritDes" and "VisOrg"
summary(tmp1)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "CritDes", ]
##
## REML criterion at convergence: 271
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.55495 -0.50027 -0.08228  0.64663  1.60935
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4349   0.6595
##  Residual             0.2473   0.4972
## Number of obs: 115, groups:  Artifact, 89
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1   1.6863     0.1207   13.98
## as.factor(Rater)2   2.1129     0.1219   17.34
## as.factor(Rater)3   1.8908     0.1219   15.51
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.244
## as.fctr(R)3 0.244  0.246
```

```
summary(tmp2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "InterpRes", ]
##
## REML criterion at convergence: 199.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5317 -0.7627  0.2635  0.6614  2.6535
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.06224  0.2495
##  Residual             0.25250  0.5025
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.70421    0.08912   30.34
## as.factor(Rater)2  2.58574    0.08912   29.01
## as.factor(Rater)3  2.13918    0.09027   23.70
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.061
## as.fctr(R)3 0.062  0.062
```

```
summary(tmp3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "VisOrg", ]
##
## REML criterion at convergence: 219.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5004 -0.3365 -0.2483  0.3841  1.8552
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2907   0.5392
##  Residual             0.1467   0.3830
## Number of obs: 115, groups:  Artifact, 89
##
## Fixed effects:
##                   Estimate Std. Error t value
## as.factor(Rater)1  2.37794    0.09658   24.62
## as.factor(Rater)2  2.64891    0.09564   27.70
## as.factor(Rater)3  2.28355    0.09658   23.64
##
## Correlation of Fixed Effects:
##             a.(R)1 a.(R)2
## as.fctr(R)2 0.263
## as.fctr(R)3 0.265  0.263
```

For rubric `CritDes`,`InterpRes` and `VisOrg`, considering the same rater, the rating for rubric is distinguishable in different artifacts. For the same artifact, rater 2 tend to give the highest rating for rubric `CritDes`, and rater 2 gives the rating that is 0.4266 higher than rater 1, and 0.2221 higher than rater 3. For the same artifact, rater 2 tend to give the highest rating for rubric `InterpRes`, and rater 1 gives the rating that is 0.11847 higher than rater 2, and 0.56503 higher than rater 3. For the same artifact, rater 2 tend to give the highest rating for rubric `VisOrg`, and rater 2 gives the rating that is 0.27097 higher than rater 1, and 0.36536 higher than rater 3.

```
fla4 <- formula(model.formula.alldata[["TxtOrg"]])
fla5 <- formula(model.formula.alldata[["RsrchQ"]])
fla6 <- formula(model.formula.alldata[["InitEDA"]])

tmp4 <- lmer(fla4,data=tall.nonmissing[tall.nonmissing$Rubric=="TxtOrg",])
tmp5 <- lmer(fla5,data=tall.nonmissing[tall.nonmissing$Rubric=="RsrchQ",])
tmp6 <- lmer(fla6,data=tall.nonmissing[tall.nonmissing$Rubric=="InitEDA",])


summary(tmp4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "TxtOrg", ]
##
```

```
## REML criterion at convergence: 247.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3557 -0.7550  0.3834  0.5302  2.4132
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09371  0.3061
##  Residual             0.39573  0.6291
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.58745    0.06821   37.93
```

```
summary(tmp5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "RsrchQ", ]
##
## REML criterion at convergence: 209.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2694 -0.5285 -0.3736  0.9743  2.4770
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07276  0.2697
##  Residual             0.27825  0.5275
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35169    0.05794   40.59
```

```
summary(tmp6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: as.numeric(Rating) ~ (1 | Artifact)
##    Data: tall.nonmissing[tall.nonmissing$Rubric == "InitEDA", ]
##
## REML criterion at convergence: 239
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8889 -0.3391 -0.1427  0.4276  1.6035
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3651   0.6042
```

```
##  Residual             0.1655   0.4068
## Number of obs: 116, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44226    0.07537    32.4
```

For rubric `TxtOrg`,`RsrchQ` and `InitEDA`, the rating for rubric is distinguishable in different artifacts.

3(iv): Trying to add fixed effects, interactions, and new random effects to the "combined" model Rating ~ 1 + (0 + Rubric|Artifact), using all the data.

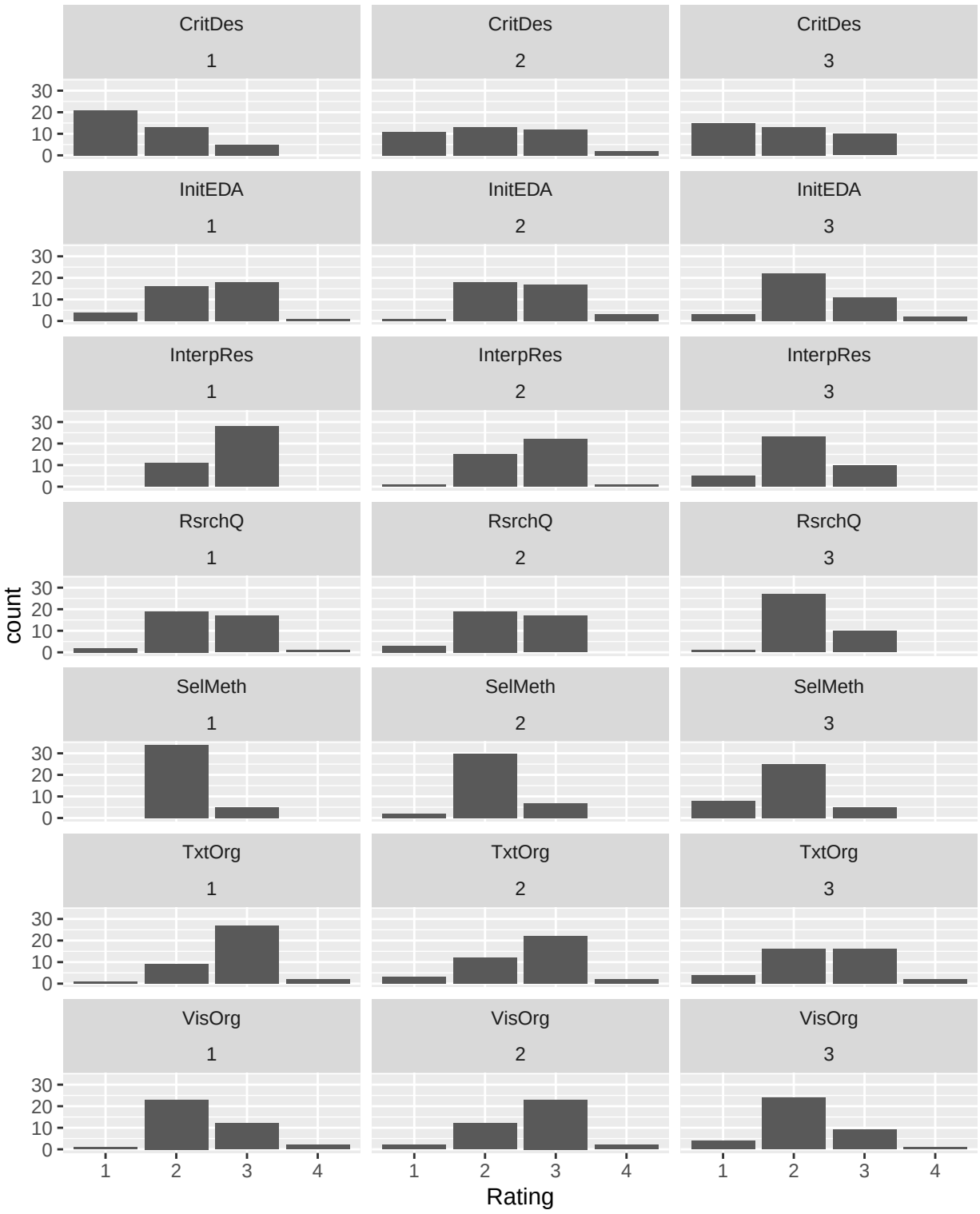
```
## Start with the "combined" intercept-only model...

comb.0 <- lmer(as.numeric(Rating) ~ 1 + (0 + Rubric | Artifact),
               data=tall.nonmissing)

summary(comb.0)


# Try adding fixed effects with no interactions...

comb.full <- update(comb.0, . ~ . + as.factor(Rater) + Semester +
                      Sex + Repeated + Rubric)

summary(comb.full)

# fixed effects selection
comb.back_elim <- fitLMER.fnc(comb.full, log.file.name = FALSE)
summary(comb.back_elim)
```

```
# try interactions
comb.inter <- update(comb.back_elim, . ~ . + as.factor(Rater)*Semester*Rubric)

ss <- getME(comb.inter,c("theta","fixef"))
comb.inter.u<- update(comb.inter,start=ss,
             control=lmerControl(optimizer="bobyqa",
                                 optCtrl=list(maxfun=2e5)))

summary(comb.inter.u)

# fixed effects interaction selection
comb.inter_elim <- fitLMER.fnc(comb.inter.u, log.file.name = FALSE)

summary(comb.inter_elim)
```

```
# the highlights for 3 models

# full model with interaction
formula(comb.inter.u)
```

```
# model after interaction selection
formula(comb.inter_elim)

# model without interaction
formula(comb.back_elim)

summary(comb.inter.u)$varcor
summary(comb.inter_elim)$varcor
summary(comb.back_elim)$varcor

anova(comb.back_elim,comb.inter_elim,comb.inter.u)
# model after interaction selection is preferred
```

```
# facets plot for each rater across all rubrics

g <- ggplot(tall.nonmissing, aes(x=Rating)) +
  geom_bar() +
  facet_wrap( ~ Rubric + Rater, nrow=7)

g
```




```
# Finally, consider adding random effects to what seems like the

## best model so far, comb.inter_elim
m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
```

```
             (0 + as.factor(Rater) | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)

anova(m0,mA)
## AIC and BIC both like including (0 + as.factor(Rater) | Artifact) in the model

m0 <- comb.inter_elim
mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + Semester | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)

anova(m0,mA)
## AIC and BIC do not like (0 + Semester | Artifact) in the model...


#m0 <- comb.inter_elim
#mA <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             #(0 + as.factor(Rater) | Artifact) +
             #(0 + as.factor(Rater):Rubric | Artifact) + as.factor(Rater) +
             #Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)
## There are not enough observations to fit mA here, so we need not do any
## formal model comparison...
```

```
# So, to summarize, the "final" model appears to be
comb.final <- lmer(as.numeric(Rating) ~ (0 + Rubric | Artifact) +
             (0 + as.factor(Rater) | Artifact) + as.factor(Rater) +
             Semester + Rubric + as.factor(Rater):Rubric, data=tall.nonmissing)

formula(comb.final)
summary(comb.final)$varcor
summary(comb.final)$coef
```

Our final model, we can interpret the pieces as follows:

(0 + as.factor(Rater) | Artifact) + as.factor(Rater) There is a kind of Rater x Artifact interaction: each Rater's rating on each Artifact differs from what we would expect (from the fixed effects alone) by a small random effect that depends on the Artifact.

(0 + Rubric | Artifact) + Rubric There is a kind of Rubric x Artifact interaction: There are different average scores on each rubric, but the rubric averages also vary a bit from one Artifact to the next, by a small random effect that depends on Artifact.

In all of this, the fact that Rubric scores depend on Artifact (that is, there is a kind of Rubric x Artifact interaction) is what we might expect: the artifacts aren't all of equal quality on each rubric, and so we should expect the average scores on each Rubric to vary from one Artifact to the next.

It does look as if the 3 raters have different ways of scoring the 7 rubrics, so the interaction we found in final model makes sense. Clearly, it is not the case that one rater is simply more harsh than another, or something like that. Among all 3 raters, for rubric `InitEDA` and `VisOrg`, rater 2 tends to give the highest score, while rater 1 tends to give the highest score for other 5 rubrics. For example, for rubric `InitEDA`, given all the other variables are the same, Rater2 rates $|0.3660743 - 0.2998406| = 0.0662337$ higher than Rater1 in semester S19. Among all rubrics, rater1 tends to give the lowest score for rubric `SelMeth`, rater 2 tends to give the lowest score for rubric `SelMeth`, rater 3 tends to give the lowest score for rubric `CritDes`. For example, for Rater1, given all the other variables are the same, he/she will rate $1.0157913 - 0.4107115 = 0.6050798$ less for rubric `SelMeth` than rubric `TxtOrg` in semester S19.

Rubric + as.factor(Rater) + as.factor(Rater):Rubric There is a Rater x Rubric interaction: each Rater uses each Rubric in a way that is not like, or even parallel to, other rater's Rubric usage, and we saw that in the facets plot above also.

More troubling are the Rater x Rubric interaction and the "kind of" Rater x Artifact interaction. The Rater x Rubric interaction suggests that the Raters are not all interpreting the Rubrics in the same way. The "kind of" Rater x Artifact interaction suggests that the Raters are not interpreting the evidence in the artifacts in the same way. These interactions suggest that perhaps the raters should be trained more, to make the raters' ratings more similar to each other.

Moreover, artifacts in semester S19 tend to get lower ratings than semester F19. For example, for any rubric rated by the same rater, given all the other variables are the same, artifacts in semester S19 will have grades 0.1591747 less than artifacts in semester F19.

Appendix 4. Anything Else Interesting about Data

– As we have mentioned above, artifacts in semester S19 tend to get lower ratings than semester F19, which may indicate the process of implementing a new "General Education" program for undergraduates is successful. However, we still need to consider other influence factors like the different difficulty level of artifacts or different experiment students.

ICC's of the final model

Next, we will compute the ICC for each rubric of the final model.

```
Rubric.names <- sort(unique(tall.nonmissing$Rubric))

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

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00493103 (tol = 0.002, component 1)
```

```
sig2 <- summary(tmp)$sigma^2
tau2 <- attr(summary(tmp)$varcor[[1]],"stddev")^2
ICC <- tau2 / (tau2 + sig2)

names(ICC) <- Rubric.names

agreement.results <- cbind(ICC.alldata=ICC.vec,ICC.final=ICC)

round(agreement.results,2)
```

```
##           ICC.alldata ICC.final
## CritDes          0.67      0.79
## InitEDA          0.69      0.70
## InterpRes        0.22      0.43
## RsrchQ           0.21      0.57
## SelMeth          0.47      0.22
## TxtOrg           0.19      0.65
## VisOrg           0.66      0.63
```

No, ICC's from these models do not agree with our earlier ICC's. For example, ICC of `CritDes` increase from 0.67 to 0.79, ICC of `InitEDA` increases from 0.69 to 0.70, ICC of `InterpRes` increase from 0.22 to 0.43, ICC of `RsrchQ` increase from 0.21 to 0.57, ICC of `SelMeth` decrease from 0.47 to 0.22, ICC of `TxtOrg` increases from 0.19 to 0.65, ICC of `VisOrg` decreases from 0.66 to 0.63.

The difference is relatively small for Rubric `VisOrg`, `InitEDA` and `CritDes`, while the difference is relatively large for Rubric `TxtOrg`, `InterpRes`, `RsrchQ` and `SelMeth`.

fall and spring semester data imbalance

```
spring <- tall[tall$Semester=="S19",]
fall <- tall[tall$Semester=="F19",]

par(mfrow=c(1,2))
barplot(table(spring$Rating), ylim = c(0,300),
        main = "Ratings for spring semester", xlab = "Rating", ylab = "Count")
barplot(table(fall$Rating), ylim = c(0,300),
        main = "Ratings for fall semester", xlab = "Rating", ylab = "Count")
```