Stefano Molina
gmolinam@andrew.cmu.edu

November 30, 2021

Abstract

The paper aims to understand the performance of students in a new program in Dietrich College. Using data from 91 artifacts developed by students and rated by three graders, a relation between information from the students, the artifacts and the graders is analyzed to explain the variability in ratings. Some visual analysis is performed to understand these relations and also statistical methods like multilevel models with variable selection are used, as well as a linear regression. The results of the multilevel model show that the ratings depend highly on a combination of both the rubrics and raters as the selected model has fixed effects for both and random effects for rubric.

Keywords: linear regression, multi-level models, hierarchical models

1 Introduction

Dietrich College at Carnegie Mellon University wants to implement a new 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. Recently the college has been experimenting with rating work in Freshman Statistics, using raters from across the college. In a recent experiment, 91 project papers—referred to as "artifacts"—were randomly sampled from a Fall and Spring section of Fresh- man Statistics. Three raters from three different departments were asked to rate these artifacts on seven rubrics, which are described in the Data Section.
The research questions for this paper are the following:

- 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?
- How are the various factors in this experiment related to the ratings? Do the factors interact in any interesting ways?
- Is there anything else interesting to say about this data?

2 Data

The data set consists of variables that contain information about the artifacts, such as information on their authors, graders, and rubrics. The last are described in the following list and the rest of them on Table 1.

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

The seven rubrics then had the following possible ratings:

- 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.

The raters were not informed about which class or student produced the artifacts. Thirteen out of 91 artifacts were graded by all three raters and the rest only one each. The rest of the variables are presented on Table 1. A final note on the data is that it was delivered in both a wide and and a long format to accommodate the analyses.

Variable	Values	Description
X	1..91	Row number in the data set
Rater	1,2,3	Which of the three raters gave a rating
Sample	1..91	Sample number
Overlap	1..13	Unique identifier for artifact seen by all 3 raters
Semester	Fall, Spring	Which semester the artifact came from
Sex	M, F	Sex or gender of student who created the artifact
RsrchQ	1,2,3,4	Rating on Research Question
CritDes	1,2,3,4	Rating on Critique Design
InitEDA	1,2,3,4	Rating on Initial EDA
SelMeth	1,2,3,4	Rating on Select Method
InterpRes	1,2,3,4	Rating on Interpret Results
VisOrg	1,2,3,4	Rating on Visual Organization
TxtOrg	1,2,3,4	Rating on Text Organization
Artifact	1..13	Unique identifier for each artifact
Repeated	0,1	1 = this is one of the 13 artifacts seen by all 3 raters

Table 1: Variable description for the data

3 Methods

3.1 Distribution of ratings

To understand the distribution of ratings for each rubric, frequency tables were created. This tables show the number of observations of every possible rating for each rating were observed. Afterwards, this quantities were used to produce bar plots for easier and faster

interpretation. The plots provide an easy to detect comparison between the distribution of ratings for each rubric and also help understand the particular patterns of each of them. Additionally, a plot for ratings grouped by rater was created. This plot will show if each rater may be more inclined to assign different grades from the others.
Finally, the same plots and tables were created for the subset of artifacts graded by all three raters.

3.2 Agreement between graders

The agreement between graders will be analyzed using two metrics: the intra-class correlation (ICC) and the percentage of agreement between graders for the same rubric.
Intra-class correlation (Gelman and Hill (2006)) is defined as the correlation between observations of the same group. This concept can be better understood defining a multilevel model as follows:

$$
\begin{aligned}
y_i &= \beta_0 + \eta_{j[i]} + \varepsilon_i \\
\eta_{j[i]} &\sim N(0, \tau^2) \\
\sigma_i &\sim N(0, \sigma^2)
\end{aligned}
$$

In this equations, $\eta_{j[i]}$ refers to the variance within groups, in this case artifacts, while σ_i refers to the variance of each rating. ICC is defined as the correlation between y_i and $y_{i'}$ if $j[i] = j[i']$ as follows: $Corr(y_i, y_{i'}) = \frac{\tau^2}{\tau^2+\sigma^2}$.

For each rubric, a multilevel model was fitted with **Artifact** as the grouping variable. Each model contains unique σ^2 and τ^2, which allow to calculate in their corresponding ICC.

The ICC will help to assess how much the ratings for each rubric resemble each other for the three raters. This statistic helps understand agreement between raters even if they do not assign the same ratings but they follow similar patterns, i.e. both raters following similar grading patterns on each rubric but with a difference of one unit.

The percentage of agreement is for each rubric and pair of graders is calculated using only the subset of artifacts graded by all raters. The number of rubrics that have the same grade for each pair of raters are summed and divided by the total rubrics rated. This is the exact percentage of agreement for the raters, which may not be representative of whether two graders agree if one of them is biased towards assigning higher or lower ratings. Together with ICC, this statistic can help understand which rubrics tend to be more controversial among the raters and how they could be focusing on different aspects of the artifacts according to the department they belong to.

3.3 Relations between factor and ratings

The first step to understand if the model that explains ratings grouping by artifact was taking the single-rubric models from Section 3.2 and performing a manual variable selection since the step-wise selection library from R was misbehaving. For each rubric, five models were used, each containing a possible fixed effect and a base model with no fixed effects. These models were compared using the *anova* command to test if adding any of the fixed effects was useful. After comparing the models and determining if any fixed effects are needed, the ICCs for each model are calculated again the same way than in Section 3.2.

As rubric can be a useful variable for the model, the data was used in a different structure that allows rubric to be used as a factor along with the other variables. The model

$Rating \sim (0 + Rubric|Artifact)$ was set as baseline model and the factors that resulted significant for any of the models from the previous paragraph were added to the model and compared against each other to test whether they add explanatory power.

Finally, the model with the selected Fixed and Random effects is tested against models with interactions for the fixed effects. This can potentially help to obtain different slopes for each Rubric and get better estimates. An anova table was used to compare the models and select the final model.

3.4 Additional remarks

As a final try to understand the relations between the data, a linear model was fitted with the data. The purpose of this was to analyze if there is still potential relations that could be possibly added to the multilevel model and understand why or why not they should be added. The model includes variable selection as well as interactions based on the findings from previous subsections.

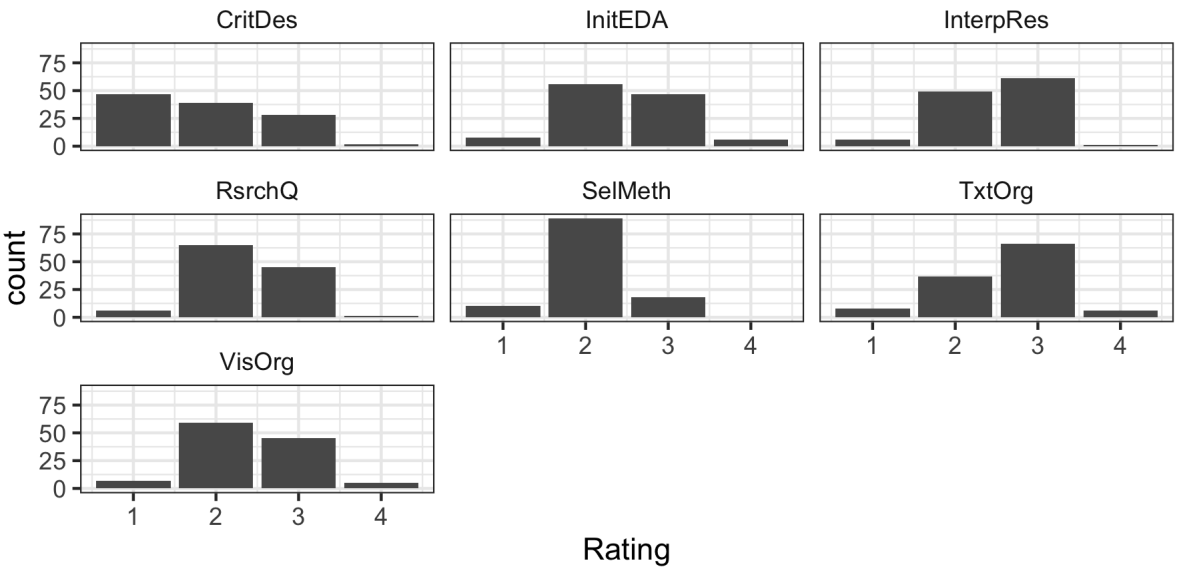
4 Results

4.1 Distribution of ratings

The distribution of ratings can be seen in Figure 1, a supplemental table with the values for this plot is presented in Appendix. A first impression is that a very low amount of rubrics get a 4, and there is even a rubric -**Selmeth**- that didn't get a single 4 grade. Other interesting result is that a high amount of 1 grades were given to the **CritDes** rubric. This results are also presented in Figure 1, where it can be seen that most of the ratings are

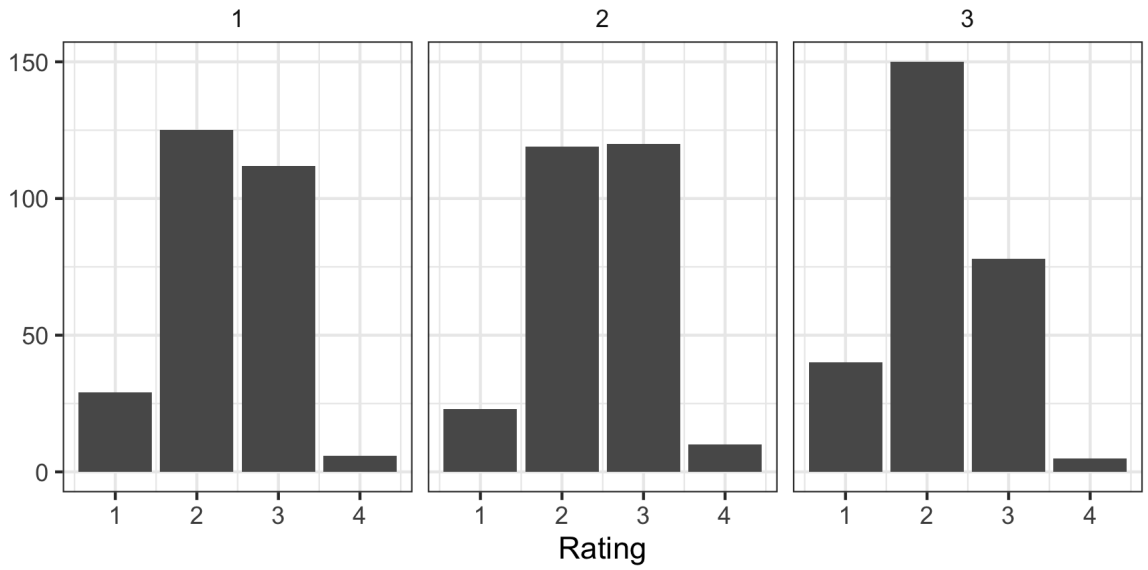
twos and threes for every rubric except **CritDes**, which has more ones.

Figure 1: Distribution of ratings for rubrics



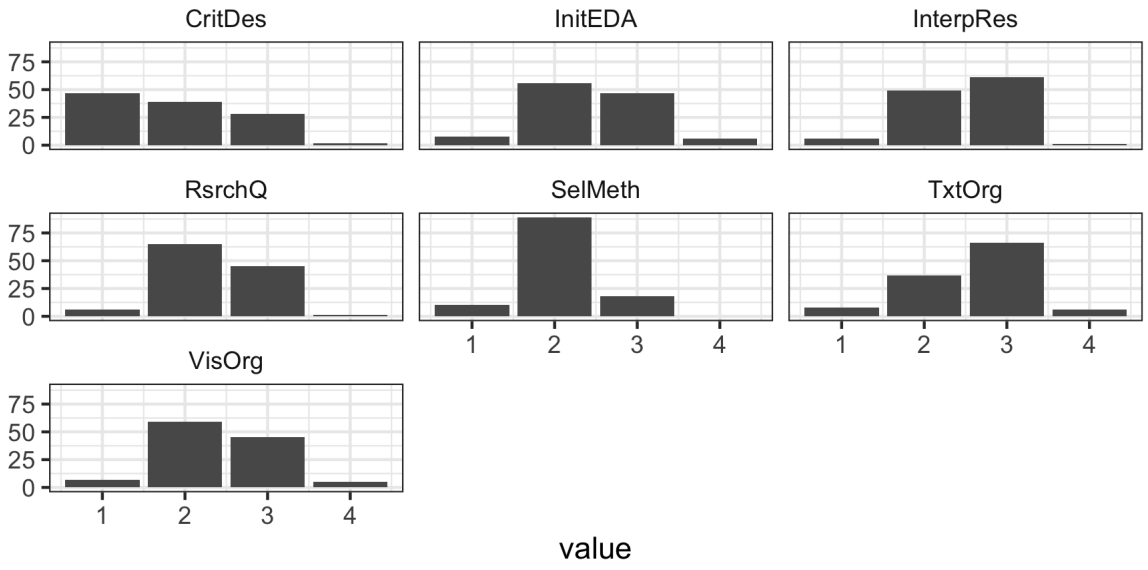
To understand how each grader assigns ratings, the distribution of ratings for each grader is included in Figure 2. The first two graders tend to have very similar distributions, having almost the same amount of ratings being 2 and 3, being low on 1 and very low on 4, whereas grader 3 tends to assign more 2 than the others. This differences will be useful for Section 4.2, when explaining the difference between ICC and percentage of agreement between raters.

Figure 2: Distribution of ratings for raters



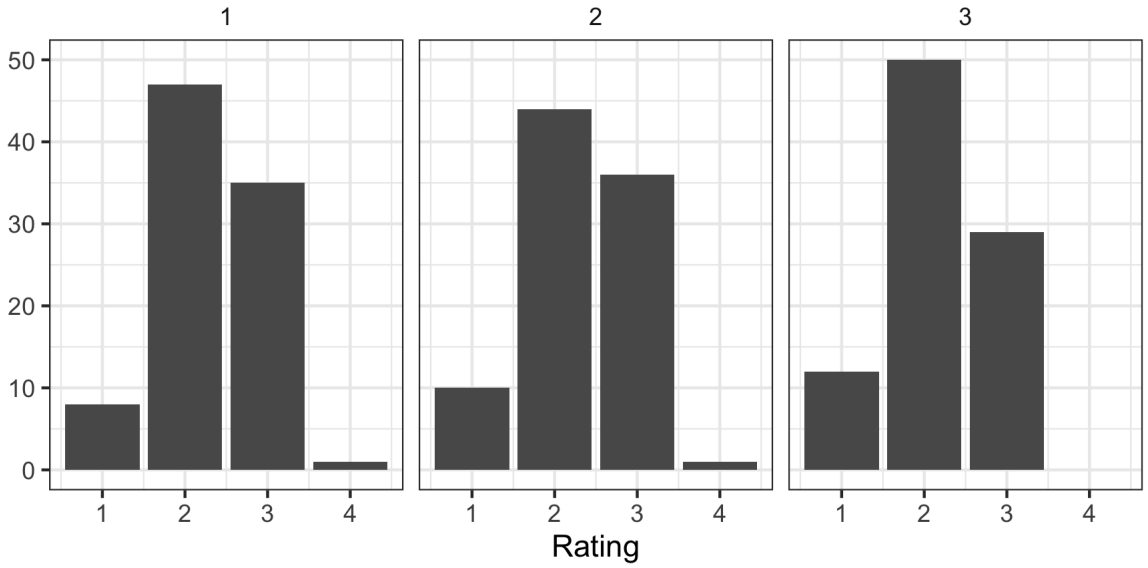
The next task was to investigate the differences between the full sample and the subsample of the 13 artifacts that were graded by the three raters. Looking at Figure 3, almost none of the artifacts that were graded by the three raters got a 4 for any rubric. This pattern may be of interest for further results and discussion. Also, the distribution of grades by rubric looks very similar to the one with all the data in Figure 1, but without most of the 4 ratings.

Figure 3: Distribution of ratings for raters



Finally, the distributions for ratings by grader in Figure 4 show that for this particular subset the graders assigned their ratings with a similar distribution across the artifacts. It is worth noting that this does not mean that they graded similarly, just that the number of ratings has a similar pattern for the three graders.

Figure 4: Distribution of ratings for raters



4.2 Agreement between raters

The ICC in Table 2 shows the correlation between the ratings for each artifact. This is done separately for every rubric to isolate the effect of each. Table 3 suggests that at least four rubrics show medium to high ICC (in the (0.49, 0.59) range), which means that graders will somehow agree in their ratings. These rubrics are: CritDes, InitEDA, SelMeth, and VisOrg. The other three rubrics have ICC below 0.22, which shows that the graders will tend to disagree when rating them.

Table 2 shows if graders tend to assign their ratings in similar patterns, but not necessarily giving the same ratings in each rubric and artifact. To find the number of ratings that are

Artifact	ICC
RsrchQ	0.19
CritDes	0.57
InitEDA	0.49
SelMeth	0.52
InterpRes	0.23
VisOrg	0.59
TxtOrg	0.14

Table 2: Intraclass correlation for rubrics

exactly the same for each grader in the rubrics, a table was created for each rubric and combination of two graders, where the combinations between ratings for each rubric and artifacts are counted. This table will have four rows and four columns, where the counts for each combinations will show and having the number of agreements in the diagonal. Adding all the elements from the diagonal and dividing by the total number of observations results in the percentage of agreement between each pair of graders. Table 3 shows the percentage of agreement for each rubric and combination of graders. The table suggests that the graders tend to agree more than half of the times for almost every rubric and that grader 1 may be in more disagreement in his ratings.

For the full dataset, i.e. including the artifacts that were revised by only one grader, the ICC was calculated again. It is shown on Table 4 This is possible because even if the data consists of artifacts that could have been graded by only one grader, there are still the ones were graded by all three. It can be seen that all the ICCs consistently change for the complete sample, which could be attributed to a change in τ^2: as the number of groups

rubric	perc_1_2	perc_1_3	perc_2_3
RsrchQ	0.38	0.77	0.54
CritDes	0.54	0.62	0.69
InitEDA	0.69	0.54	0.85
SelMeth	0.92	0.62	0.69
InterpRes	0.62	0.54	0.62
VisOrg	0.54	0.77	0.77
TxtOrg	0.69	0.62	0.54

Table 3: Percentage of agreement between graders for rubric

-artifacts in this case- the variance may change thus making the ICC change. It is not possible to construct the percentage of agreement table because there is no way to compare directly how the graders agree or disagree on a particular rating for an artifact.

4.3 Relations between factors

Table 5 shows the best model for each rubric. Three of the models suggest that using the Rater Fixed Effect could be beneficial. Also, one of the models suggests using the Sex fixed effect, which at least for now will be left out since it's only on one model. The rater fixed effect would make a lot of sense to be having some kind of influence on the rating as it is presented on the distribution of ratings by grader figure. If a grader is inclined to assign lower ratings compared to the others, it would be expected that the model can capture that behavior. The next step consists of using the Rater Fixed Effect model to construct new ICCs and compare to the previous. As the output shows, the same three models for the rubrics that suggested including th eRater Fixed Effect, which are TxtOrg, InterpRes,

Artifact	ICC
RsrchQ	0.21
CritDes	0.67
InitEDA	0.69
SelMeth	0.47
InterpRes	0.22
VisOrg	0.66
TxtOrg	0.19

Table 4: ICC for the full dataset

and SelMeth have a significant coefficient for it.

The ICCs for the model with Rater fixed effects have some effect on the original ICCs with some increases and decreases in them but no big changes on their magnitudes. This could suggest that including the Rater fixed effect is just helping get a better estimation for the ICC but the original models already did a good job.

For the model with the rubric fixed effects, the results show that adding additional either fixed or random effects makes the BIC greater, which means that they do not add explanatory power to the model with no random effects besides Rubric. This model only needs to be tested if it could use an interaction between the Rater and Rubric Fixed Effects.

Table for final model in progress

The final model is $Rating \sim Rater + Rubric + Rubric * Rater + (1 + Rubric|Artifact)$,

Table 5

	Rubric:						
	RsrchQ	InitEDA	CritDes	SelMeth	InterpRes	TxtOrg	VisOrg
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SexF				-1.027^{**}			
				(0.476)			
SexM				-0.826^{*}			
				(0.477)			
Rater					-0.272^{***}	-0.161^{**}	
					(0.062)	(0.075)	
Constant	2.358^{***}	2.448^{***}	1.907^{***}	3.000^{***}	3.029^{***}	2.914^{***}	2.445^{***}
	(0.058)	(0.075)	(0.089)	(0.471)	(0.135)	(0.164)	(0.071)
Observations	117	117	116	117	117	117	116
Log Likelihood	−105.533	−120.388	−138.935	−76.061	−101.752	−123.913	−113.209
Akaike Inf. Crit.	217.066	246.776	283.869	162.121	211.504	255.826	232.417
Bayesian Inf. Crit.	225.352	255.063	292.130	175.932	222.553	266.875	240.678

Note: *p<0.1; **p<0.05; ***p<0.01

a random intercept and random slope with group level predictors. It looks like Rater and Rubric are the only factors that have strong influence in the ratings but also the combination of both. Since all artifacts have the same rubrics, it would be expected that they do not play a important role in the model, but when taken into account with rater, they become important if the raters are more inclined to assign certain ratings to rubrics, as shown in Q1. The interaction between rater and rubric shows that these holds true for the selected model, where most of the interactions are significant (for the purpose of this paper, a coefficient is considered significant if the absolute value of their t-value is greater than two). It is also interesting to note that adding the interaction takes significance off rater, meaning that the rater alone does not have influence on the rating but the patterns that each rater show for the rubrics do.

4.4 Additional remarks

Using a regular linear model helps understand the influence of the explanatory variables in the outcome of the ratings. It is worth noting that all of the variables except for **Repeated** have significant coefficients, which means that at some point all of them can help explain the variation of **Ratings**. Also, the R^2 is high and the RSS is low. The diagnostic plots interpretation is tricky, since there are only four possible values in the response variable. This regression would be closer to a multinomial regression and the Residuals vs Fitted and Scale-Location plots shows some patterns that are not useful for interpretation, but the QQ plot and the Residuals vs Leverage plots suggest that it may be a good fit with no high influence points. This can also occur because there are no variables that can actually show outliers and most variables are factors.

Table 6

	Estimate	Std. Error	t value	Pr(>|t|)
Rater1	2.436	0.267	9.112	0
Rater2	2.974	0.268	11.103	0
Rater3	2.722	0.261	10.408	0
SemesterS19	-0.129	0.052	-2.499	0.013
SexF	-0.794	0.248	-3.195	0.001
SexM	-0.825	0.248	-3.326	0.001
RubricInitEDA	0.821	0.146	5.625	0.00000
RubricInterpRes	1.128	0.146	7.735	0
RubricRsrchQ	0.846	0.146	5.801	0
RubricSelMeth	0.538	0.146	3.692	0.0002
RubricTxtOrg	1.179	0.146	8.087	0
RubricVisOrg	0.806	0.147	5.492	0.00000
Rater2:RubricInitEDA	-0.386	0.207	-1.865	0.063
Rater3:RubricInitEDA	-0.385	0.206	-1.865	0.063
Rater2:RubricInterpRes	-0.668	0.207	-3.228	0.001
Rater3:RubricInterpRes	-0.872	0.206	-4.226	0.00003
Rater2:RubricRsrchQ	-0.617	0.207	-2.980	0.003
Rater3:RubricRsrchQ	-0.487	0.206	-2.362	0.018
Rater2:RubricSelMeth	-0.540	0.207	-2.608	0.009
Rater3:RubricSelMeth	-0.487	0.206	-2.362	0.018
Rater2:RubricTxtOrg	-0.719	0.207	-3.475	0.001
Rater3:RubricTxtOrg	-0.641	0.206	-3.108	0.002
Rater2:RubricVisOrg	-0.295	0.208	-1.420	0.156
Rater3:RubricVisOrg	-0.499	0.207	-2.410	0.016

5 Discussion

5.1 Distribution of ratings

According to the results shown in Section 4.1, all of the rubrics have different distributions. This suggests that although graders can rate differently, the students don't appear to have the same skills for every rubric, being the Interpretation of Results and Text Organization the ones that are most developed. On the other hand, their Critique Design skills are still not well developed as they may require a lot more practice than what they have. The other four rubrics look like they have similar distribution in terms of having mostly 2 and 3 ratings, but it is worth noting that all four have modes equal to 2, which means that students generate evidence with significant flaws. This interpretation holds for the subset of data of the artifacts rated by all three graders.

The distribution of ratings for the raters are similar for raters 1 and 2, having mostly 2 and 3 ratings in similar quantities. This would be expected given the short number of possible ratings, having most students being competent enough or a little off track and not many outstanding observations in either way. Moreover, rater number 3 tends to assign more 2 ratings, which could mean either than he is a more harsh grader or understands best the content of the artifacts. This can have a big influence in the outcome of the artifacts that were only graded by this rater, having lower but possibly more authentic final ratings.

5.2 Agreement between graders

Some rubrics appear to cause more agreement between raters than others. As mentioned in Section 4.2, four rubrics (CritDes, InitEDA, SelMeth, and VisOrg) have high ICC, which

mean that raters tend to agree at some degree in the ratings they assign to them. The other three rubrics usually refer to skills that students will develop during their university years and may cause controversy depending on the graders' background. As Section 4.1 suggests, one of the graders tends to assign lower grades overall, which could be having big influence on the outcome of ICC. Furthermore, if indeed this rater is an expert on the subject of the artifact, it is to be expected that he/she will give more attention to the RsrchQ and InterpRes rubrics since they are the motive of the artifact.

The percentage of agreement table helps to understand how much the raters tend to agree on each rubric. Looking at the table, there may be some slight pattern showing that rater 1 assigns grades differently from the other two, but given the sample size of 13 this is not confirmatory. Overall, all three raters agree more than 50 percent of the time for all but one rubric: RsrchQ is below this percentage of raters 1 and 2. This agrees with the statement in the last paragraph where one rater could be assigning lower ratings at some rubrics.

5.3 Relations between factors

Work in progress.

5.4 Additional remarks

Using this model can help for further MLM analyses with the current data, but also understand the differences between the variables is useful to potentially help the MLM model point in the correct direction as some of the commands depend on Convex Optimization and fail from time to time. Knowing that most of the variables could help the model suggests that the study could take a step back to analyze if any of them could be added to

the MLM model's fixed effects or why not.

References

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Appendix

1

The data consists of two datasets with the same information but presented in wide and long formats. The wide format contains variables for: Rater, Sample, Overlap, Semester, Sex, the seven rubrics, artifact and Repeated. The last variable is just an indicator that tells if an artifact was graded by all three graders.

To better understand how the grades for each rubric were given, a frequency table for each rubric and grade was generated, a first impression is that a very low amount of rubrics get a 4, and there is even a rubric - **Selmeth**- that didn't get a single 4 grade. Other interesting result is that a high amount of 1 grades were given to the **CritDes** rubric. This results are also presented in Figure 1, where it can be seen that most of the ratings are twos and threes for every rubric except **CritDes**, which has more ones. Table 2 includes mean, standard deviation, and median for each rubric but none of them seem to add something useful to the information from Table 1.

```
ratings <- read.csv("/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/ratings.csv")
```

```
tall <- read.csv("/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/tall.csv")
```

```
freqs <- bind_rows(table(ratings$RsrchQ),
                   table(ratings$CritDes),
                   table(ratings$InitEDA),
                   table(ratings$SelMeth),
                   table(ratings$InterpRes),
                   table(ratings$VisOrg),
                   table(ratings$TxtOrg))

freqs$name <- c("RsrchQ", "CritDes", "InitEDA", "SelMeth", "InterpRes", "VisOrg", "TxtOrg")

freqs <- freqs %>% dplyr::select(name, everything())
```

```
knitr::kable(freqs, caption = "Frequency of ratings by rubric")
```

Table 1: Frequency of ratings by rubric

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

```
dist <- ratings %>%
  dplyr::select(X, RsrchQ, CritDes, InitEDA, SelMeth, InterpRes, VisOrg, TxtOrg) %>%
  pivot_longer(!X)

dist_table <- data.table(dist)[,
                               .(Mean = mean(value, na.rm = T),
```

```
                                  SD = sd(value, na.rm = T),
                                  Median = median(value, na.rm = T)),
                                by = "name"]
```

```
xtable(dist_table)
```

```
knitr::kable(dist_table, caption = "Mean, Standard Deviation and median ratings by rubric")
```

Table 2: Mean, Standard Deviation and median ratings by rubric

name	Mean	SD	Median
RsrchQ	2.350427	0.5918446	2
CritDes	1.870690	0.8395669	2
InitEDA	2.435897	0.6995641	2
SelMeth	2.068376	0.4864810	2
InterpRes	2.487179	0.6104744	3
VisOrg	2.413793	0.6733300	2
TxtOrg	2.598291	0.6955503	3

```
fig1 <- dist %>% ggplot()+
  geom_bar(aes(value))+
  facet_wrap(~name)+
  theme_bw()+
  xlab("Rating")+
  theme(strip.background = element_blank())

fig1
```

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

```
ggsave(plot = fig1, "/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/Plots/fig1.png",
       width = 6,
       height = 3,
       dpi = 300)
```

Additionally, to understand how each grader assigns ratings, the distribution of ratings for each grader is included in Figure 2. The first two graders tend to have very similar distributions, having almost the same amount of ratings being 2 and 3, being low on 1 and very low on 4, whereas grader 3 tends to assign more 2 than the others.

```
fig2 <- tall %>%
  group_by(Rater, Rating) %>%
  dplyr::summarise(n = n()) %>%
  ggplot() +
  geom_col(aes(Rating,n))+
  facet_grid(~Rater)+
  theme_bw()+
  ylab("")+
  theme(strip.background = element_blank())
```

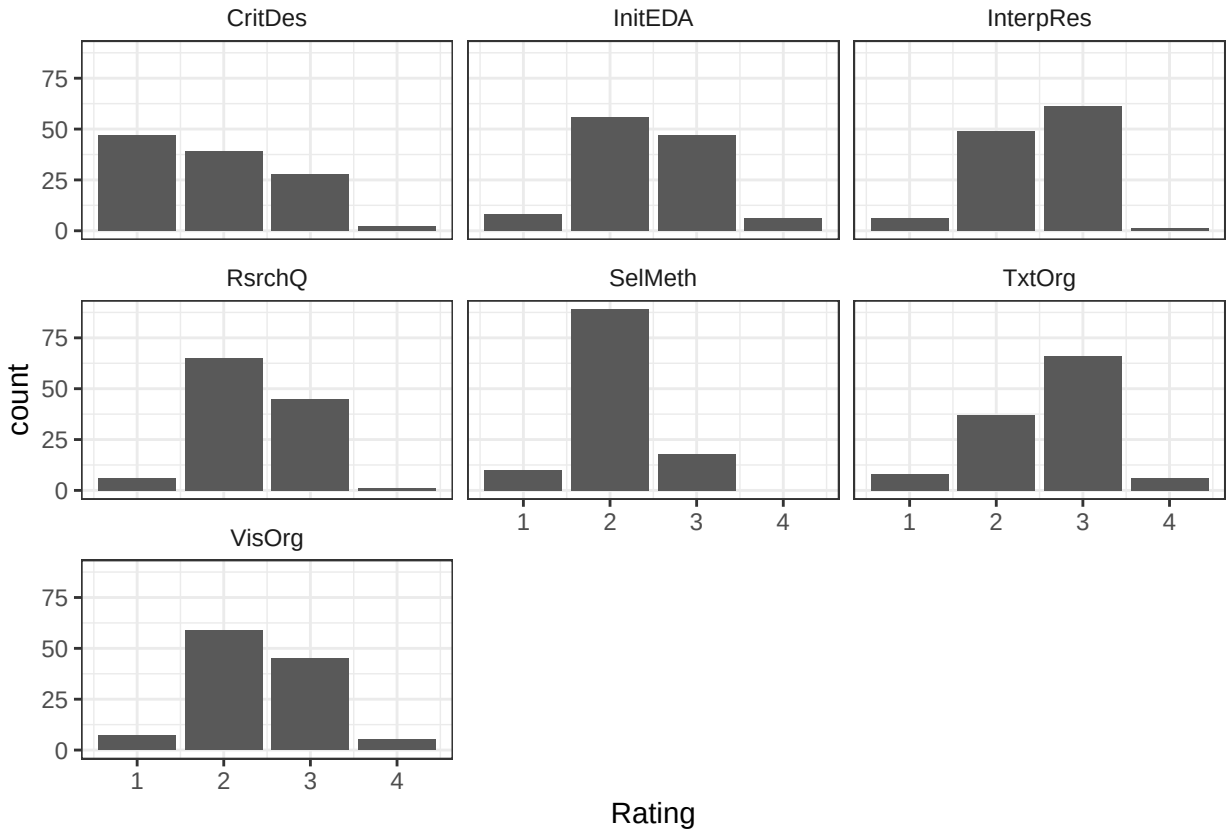


Figure 1: Distribution of ratings by rubric

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

```
fig2
```

```
## Warning: Removed 2 rows containing missing values (position_stack).
```

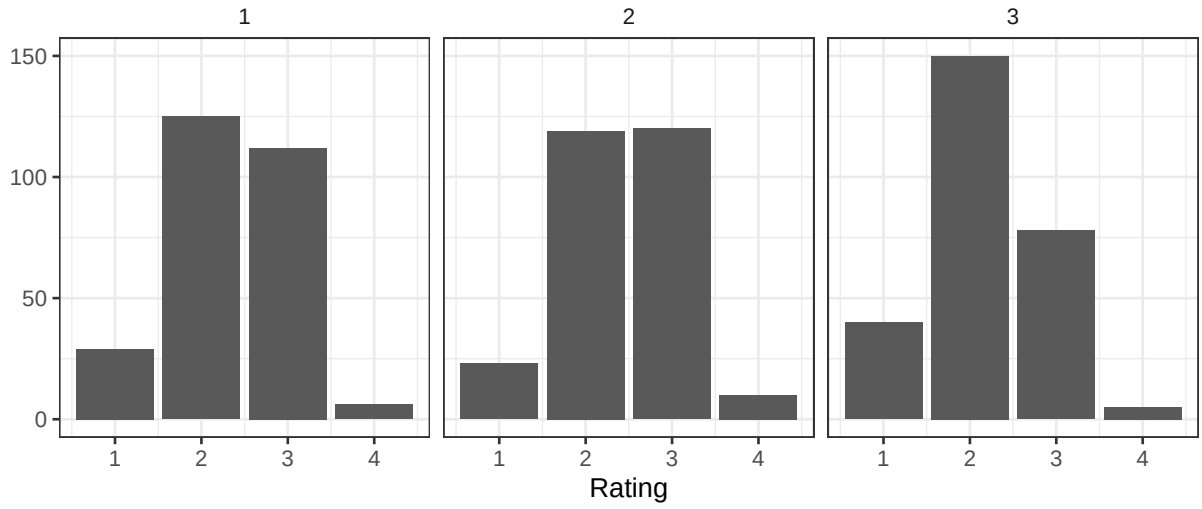


Figure 2: Distribution of ratings by grader

```
ggsave(plot = fig2, "/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/Plots/fig2.png",
       width = 6,
       height = 3,
       dpi = 300
       )
```

Now, subsetting for the 13 artifacts seen by the three graders the same statistics and plot are performed. Looking at both Table 3 and Figure 3, almost none of the artifacts that were graded by the three raters got a 4 for any rubric. This pattern may be of interest for further results and discussion. Also, the distribution of grades by rubric looks very similar to the one with all the data, but without most of the 4 ratings.

```
ratings_subset <- ratings %>% filter(Repeated == 1)
freqs <- bind_rows(table(ratings_subset$RsrchQ),
                   table(ratings_subset$CritDes),
                   table(ratings_subset$InitEDA),
                   table(ratings_subset$SelMeth),
                   table(ratings_subset$InterpRes),
                   table(ratings_subset$VisOrg),
                   table(ratings_subset$TxtOrg))

freqs$name <- c("RsrchQ", "CritDes", "InitEDA", "SelMeth", "InterpRes", "VisOrg", "TxtOrg")

freqs <- freqs %>% dplyr::select(name, everything())
```

Table 3: Frequency of ratings by rubric

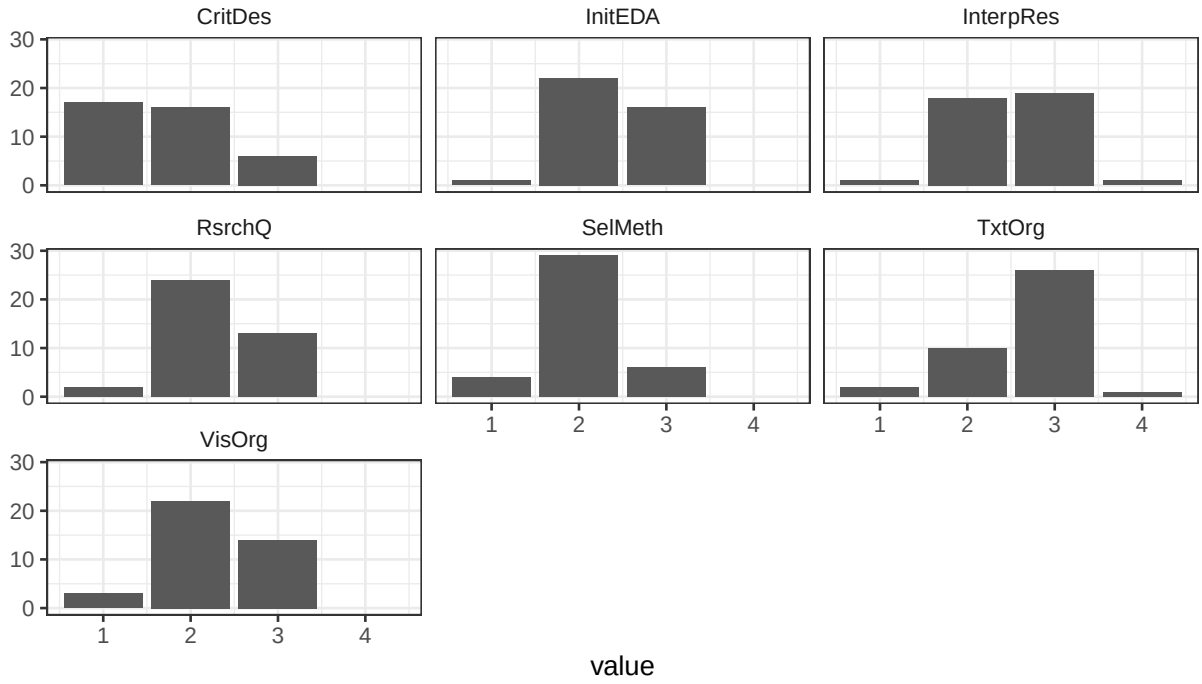
name	1	2	3	4
RsrchQ	2	24	13	NA
CritDes	17	16	6	NA
InitEDA	1	22	16	NA
SelMeth	4	29	6	NA
InterpRes	1	18	19	1
VisOrg	3	22	14	NA
TxtOrg	2	10	26	1

```
dist <- ratings %>%
  filter(Repeated == 1) %>%
  dplyr::select(X, RsrchQ, CritDes, InitEDA, SelMeth, InterpRes, VisOrg, TxtOrg) %>%
  pivot_longer(!X)

dist_table <- data.table(dist)[,
                               .(Mean = mean(value, na.rm = T),
                                 SD = sd(value, na.rm = T),
                                 Median = median(value, na.rm = T)),
                               by = "name"]
```

Table 4: Mean, Standard Deviation and median ratings by rubric

name	Mean	SD	Median
RsrchQ	2.282051	0.5595448	2
CritDes	1.717949	0.7236137	2
InitEDA	2.384615	0.5436419	2
SelMeth	2.051282	0.5103517	2
InterpRes	2.512821	0.6013929	3
VisOrg	2.282051	0.6047495	2
TxtOrg	2.666667	0.6212607	3

```
fig3 <- dist %>% ggplot()+
  geom_bar(aes(value))+
  facet_wrap(~name)+
  theme_bw()+
  ylab("")+
  theme(strip.background = element_blank())

fig3
```




```
ggsave(plot = fig3, "/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/Plots/fig3.png",
       width = 6,
       height = 3,
       dpi = 300
       )
```

The distributions for ratings by grader show that for this particular subset the graders assigned their ratings in a similar amount to the artifacts. It is worth noting that this does not mean that they graded similarly, just that the number of ratings has a similar pattern for the three graders.

```
fig4 <- tall %>%
  filter(Repeated == 1) %>%
  group_by(Rater, Rating) %>%
  dplyr::summarise(n = n()) %>%
  ggplot() +
  geom_col(aes(Rating,n))+
  facet_grid(~Rater)+
  theme_bw()+
  ylab("")+
  theme(strip.background = element_blank())
```

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

```
fig4
```

```
ggsave(plot = fig4, "/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/Plots/fig4.png",
       width = 6,
       height = 3,
       dpi = 300
       )
```

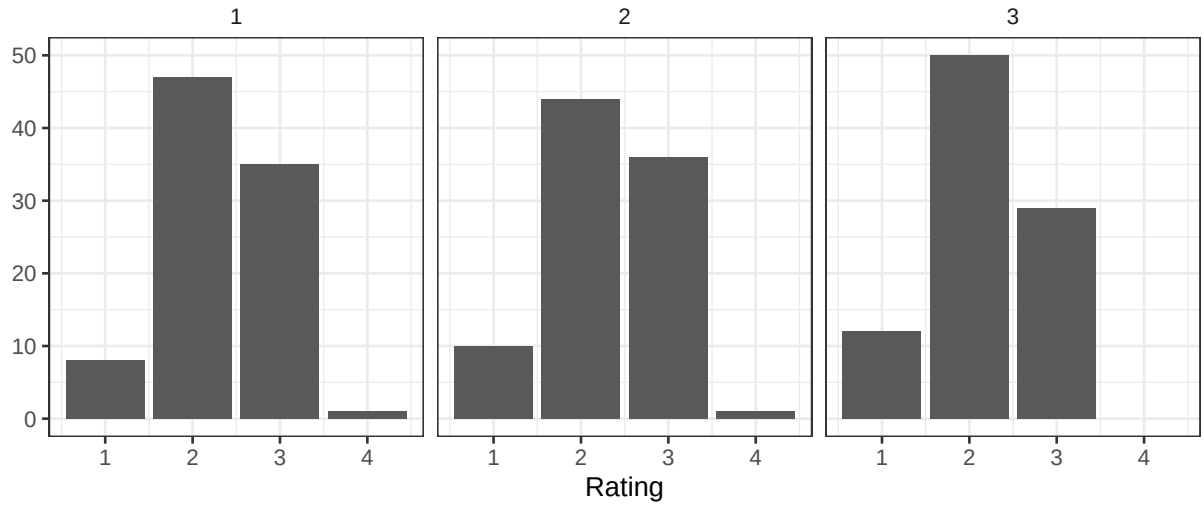


Figure 3: Distribution of ratings by grader

2

To answer if the raters tend to agree in their scores, the ICC (intra-class correlation) was calculated for each rubric. The ICC in Table 5 shows the correlation between the ratings for each artifact. This is done separately for every rubric to isolate the effect of each. Table 5 suggests that at least four rubrics show medium to high ICC (in the (0.49, 0.59) range), which means that graders will somehow agree in their ratings. These rubrics are: **CritDes**, **InitEDA**, **SelMeth**, and **VisOrg**. The other three rubrics have ICC below 0.22, which shows that the graders will tend to disagree when rating them.

```
icc_df <- rep(NA, 7)
j = 1
for(i in c("RsrchQ", "CritDes", "InitEDA", "SelMeth", "InterpRes", "VisOrg", "TxtOrg")){
  form <- formula(paste0(i, "~ 1 + (1 | Artifact)"))
  rnd_int <- summary(lmer(form, data = ratings_subset))
  icc_df[j] <- rnd_int$varcor$Artifact/(rnd_int$varcor$Artifact + rnd_int$sigma^2)
  j = j+1
}

icc_df <- data.frame(Artifact = c("RsrchQ", "CritDes",
                                  "InitEDA", "SelMeth",
                                  "InterpRes", "VisOrg",
                                  "TxtOrg"),
                     ICC = icc_df)
```

Table 5: ICC for the subset data

Artifact	ICC
RsrchQ	0.1891892
CritDes	0.5725594
InitEDA	0.4929577
SelMeth	0.5212766
InterpRes	0.2295720

Artifact	ICC
VisOrg	0.5924529
TxtOrg	0.1428571

```
print(xtable(icc_df), include.rownames=FALSE)
```

The previous table shows if graders tend to assign their ratings in similar patterns, but not necessarily giving the same ratings in each rubric and artifact. To find the number of ratings that are exactly the same for each grader in the rubrics, a table was created for each rubric and combination of two graders, where the combinations between ratings for each rubric and artifacts are counted. This table will have four rows and four columns, where the counts for each combinations will show and having the number of agreements in the diagonal. Adding all the elements from the diagonal and dividing by the total number of observations results in the percentage of agreement between each pair of graders. Table 6 shows the percentage of agreement for each rubric and combination of graders. The table suggests that the graders tend to agree more than half of the times for almost every rubric and that grader 1 may be in more disagreement in his ratings.

```
concordance <- data.frame(rubric = c("RsrchQ", "CritDes",
                                     "InitEDA", "SelMeth",
                                     "InterpRes", "VisOrg",
                                     "TxtOrg"),
                          perc_1_2 = rep(NA, 7),
                          perc_1_3 = rep(NA, 7),
                          perc_2_3 = rep(NA, 7))

j <- 1
for(rubric in c("RsrchQ", "CritDes",
                "InitEDA", "SelMeth",
                "InterpRes", "VisOrg",
                "TxtOrg")){
  i <- 1
  for(grader in 3:1){
    ratings_table <- ratings_subset %>%
      filter(Rater != grader)
    ratings_table <- ratings_table[,c("Rater", rubric, "Artifact")]
    names(ratings_table) <- c("Rater", "Rubric", "Artifact")
    ratings_table <- ratings_table %>%
      pivot_wider(names_from = Rater, values_from = Rubric)

    ratings_table[[2]] <- factor(ratings_table[[2]], levels = 1:4)
    ratings_table[[3]] <- factor(ratings_table[[3]], levels = 1:4)

    concordance[j,i+1] <- sum(diag(prop.table(table(ratings_table[[2]],
                                                    ratings_table[[3]]))))

    i <- i+1
  }
  j <- j+1
}
```

Table 6: Proportion of questions that each combination of graders agree with for every rubric

rubric	perc_1_2	perc_1_3	perc_2_3
RsrchQ	0.3846	0.7692	0.5385
CritDes	0.5385	0.6154	0.6923
InitEDA	0.6923	0.5385	0.8462
SelMeth	0.9231	0.6154	0.6923
InterpRes	0.6154	0.5385	0.6154
VisOrg	0.5385	0.7692	0.7692
TxtOrg	0.6923	0.6154	0.5385

```
print(xtable(concordance), include.rownames = F)
```

For the full dataset, i.e. including the artifacts that were revised by only one grader, the ICC was calculated again. This is possible because even if the data consists of artifacts that could have been graded by only one grader, there are still the ones were graded by all three. It can be seen that all the ICCs consistently change for the complete sample, which could be attributed to a change in τ^2: as the number of groups -artifacts in this case- the variance may change thus making the ICC change.

It is not possible to construct the percentage of agreement table because there is no way to compare directly how the graders agree or disagree on a particular rating for an artifact.

```
icc_df_all <- rep(NA, 7)
j = 1
for(i in c("RsrchQ", "CritDes",
           "InitEDA", "SelMeth",
           "InterpRes", "VisOrg",
           "TxtOrg")){
  form <- formula(paste0(i, "~ 1 + (1 | Artifact)"))
  rnd_int <- summary(lmer(form, data = ratings))
  icc_df_all[j] <- rnd_int$varcor$Artifact/(rnd_int$varcor$Artifact + rnd_int$sigma^2)
  j = j+1
}

icc_df_all <- data.frame(Artifact = c("RsrchQ", "CritDes",
                                      "InitEDA", "SelMeth",
                                      "InterpRes", "VisOrg",
                                      "TxtOrg"),
                         ICC = icc_df_all)
```

Table 7: ICC for the complete data

Artifact	ICC
RsrchQ	0.2096214
CritDes	0.6730647
InitEDA	0.6867210
SelMeth	0.4719014
InterpRes	0.2200285
VisOrg	0.6607372
TxtOrg	0.1879927

```
print(xtable(icc_df_all), include.rownames = F)
```

3

To test if additional fixed effects should be added to the model, a variable selection will be performed for each individual model on rubrics. Initially, the **lmer()** command was intended to be used, but since it was misbehaving a manual variable selection was tried. For each rubric, five models were used, each contaning a possible fixed effect and a base model with no fixed effects. These models were compared using the **anova()** command to test if adding any of the fixed effects was useful.

```
i <- "RsrchQ"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

## Data: ratings
## Models:
## rnd_int0: RsrchQ ~ 1 + (1 | Artifact)
## rnd_int1: RsrchQ ~ 1 + Rater + (1 | Artifact)
## rnd_int2: RsrchQ ~ 1 + Semester + (1 | Artifact)
## rnd_int4: RsrchQ ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: RsrchQ ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 213.19 221.48 -103.60   207.19
## rnd_int1    4 213.39 224.44 -102.70   205.39 1.8008  1     0.1796
## rnd_int2    4 214.57 225.62 -103.28   206.57 0.0000  0
## rnd_int4    4 214.57 225.62 -103.28   206.57 0.0017  0
## rnd_int3    5 215.37 229.18 -102.68   205.37 1.1983  1     0.2737
```

```
rnd_rsrchq <- rnd_int0
```

```
i <- "CritDes"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

```
## Data: ratings
## Models:
## rnd_int0: CritDes ~ 1 + (1 | Artifact)
## rnd_int1: CritDes ~ 1 + Rater + (1 | Artifact)
## rnd_int2: CritDes ~ 1 + Semester + (1 | Artifact)
## rnd_int4: CritDes ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: CritDes ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 280.86 289.12 -137.43   274.86
## rnd_int1    4 280.76 291.77 -136.38   272.76 2.0985  1     0.1474
## rnd_int2    4 282.58 293.60 -137.29   274.58 0.0000  0
## rnd_int4    4 281.85 292.87 -136.93   273.85 0.7294  0
## rnd_int3    5 282.65 296.42 -136.33   272.65 1.1972  1     0.2739
```

```
rnd_critdes <- rnd_int0
```

```
i <- "InitEDA"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

```
## Data: ratings
## Models:
## rnd_int0: InitEDA ~ 1 + (1 | Artifact)
## rnd_int1: InitEDA ~ 1 + Rater + (1 | Artifact)
## rnd_int2: InitEDA ~ 1 + Semester + (1 | Artifact)
## rnd_int4: InitEDA ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: InitEDA ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 243.42 251.71 -118.71   237.42
## rnd_int1    4 243.26 254.31 -117.63   235.26 2.1635  1     0.1413
## rnd_int2    4 245.38 256.43 -118.69   237.38 0.0000  0
## rnd_int4    4 245.27 256.32 -118.63   237.27 0.1153  0
## rnd_int3    5 246.75 260.56 -118.38   236.75 0.5174  1     0.4720
```

```
rnd_initeda <- rnd_int0
```

```
i <- "SelMeth"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
```

```
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

## Data: ratings
## Models:
## rnd_int0: SelMeth ~ 1 + (1 | Artifact)
## rnd_int1: SelMeth ~ 1 + Rater + (1 | Artifact)
## rnd_int2: SelMeth ~ 1 + Semester + (1 | Artifact)
## rnd_int4: SelMeth ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: SelMeth ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 159.53 167.82 -76.768   153.53
## rnd_int1    4 157.43 168.48 -74.714   149.43 4.1064  1   0.042721 *
## rnd_int2    4 148.64 159.69 -70.322   140.64 8.7848  0
## rnd_int4    4 161.49 172.54 -76.745   153.49 0.0000  0
## rnd_int3    5 155.32 169.13 -72.660   145.32 8.1702  1   0.004258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rnd_selmeth <- rnd_int3
```

```
i <- "InterpRes"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

## Data: ratings
## Models:
## rnd_int0: InterpRes ~ 1 + (1 | Artifact)
## rnd_int1: InterpRes ~ 1 + Rater + (1 | Artifact)
## rnd_int2: InterpRes ~ 1 + Semester + (1 | Artifact)
## rnd_int4: InterpRes ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: InterpRes ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC   logLik deviance   Chisq Df Pr(>Chisq)
## rnd_int0    3 220.09 228.38 -107.048   214.09
```

```
## rnd_int1    4 203.79 214.84  -97.897   195.79 18.3021  1  1.885e-05 ***
## rnd_int2    4 221.76 232.81 -106.878   213.76  0.0000  0
## rnd_int4    4 222.01 233.06 -107.007   214.01  0.0000  0
## rnd_int3    5 223.14 236.95 -106.572   213.14  0.8708  1     0.3507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rnd_interpres <- rnd_int1
```

```
i <- "VisOrg"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater + (1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

```
## Data: ratings
## Models:
## rnd_int0: VisOrg ~ 1 + (1 | Artifact)
## rnd_int1: VisOrg ~ 1 + Rater + (1 | Artifact)
## rnd_int2: VisOrg ~ 1 + Semester + (1 | Artifact)
## rnd_int4: VisOrg ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: VisOrg ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 228.95 237.21 -111.47   222.95
## rnd_int1    4 230.40 241.42 -111.20   222.40 0.5461  1     0.4599
## rnd_int2    4 229.33 240.34 -110.67   221.33 1.0735  0
## rnd_int4    4 229.76 240.77 -110.88   221.76 0.0000  0
## rnd_int3    5 231.47 245.23 -110.73   221.47 0.2937  1     0.5879
```

```
rnd_visorg <- rnd_int0
```

```
i <- "TxtOrg"
form <- formula(paste0(i, "~ 1 +(1 | Artifact)"))
rnd_int0 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Rater +(1 | Artifact)"))
rnd_int1 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 +Semester +(1 | Artifact)"))
rnd_int2 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Sex  + (1 | Artifact)"))
rnd_int3 <- lmer(form, data = ratings)
form <- formula(paste0(i, "~ 1 + Repeated + (1 | Artifact)"))
rnd_int4 <- lmer(form, data = ratings)
anova(rnd_int0, rnd_int1, rnd_int2, rnd_int3, rnd_int4)
```

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

## Data: ratings
## Models:
## rnd_int0: TxtOrg ~ 1 + (1 | Artifact)
## rnd_int1: TxtOrg ~ 1 + Rater + (1 | Artifact)
## rnd_int2: TxtOrg ~ 1 + Semester + (1 | Artifact)
## rnd_int4: TxtOrg ~ 1 + Repeated + (1 | Artifact)
## rnd_int3: TxtOrg ~ 1 + Sex + (1 | Artifact)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## rnd_int0    3 251.45 259.74 -122.73   245.45
## rnd_int1    4 248.88 259.93 -120.44   240.88 4.5725  1    0.03249 *
## rnd_int2    4 251.92 262.97 -121.96   243.92 0.0000  0
## rnd_int4    4 252.99 264.04 -122.49   244.99 0.0000  0
## rnd_int3    5 254.99 268.80 -122.50   244.99 0.0000  1    1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rnd_txtorg <- rnd_int1
```

```
stargazer::stargazer(rnd_rsrchq, rnd_initeda, rnd_critdes, rnd_selmeth, rnd_interpres, rnd_txtorg, rnd_
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Mon, Nov 29, 2021 - 19:14:24
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{7}{c}{\textit{Dependent variable:}} \\
## \cline{2-8}
## \\[-1.8ex] & RsrchQ & InitEDA & CritDes & SelMeth & InterpRes & TxtOrg & VisOrg \\
## \\[-1.8ex] & (1) & (2) & (3) & (4) & (5) & (6) & (7)\\
## \hline \\[-1.8ex]
##  SexF &  &  &  & $-$1.027$^{**}$ &  &  &  \\
##   &  &  &  & (0.476) &  &  &  \\
##   & & & & & & & \\
##  SexM &  &  &  & $-$0.826$^{*}$ &  &  &  \\
##   &  &  &  & (0.477) &  &  &  \\
##   & & & & & & & \\
##  Rater &  &  &  &  & $-$0.272$^{***}$ & $-$0.161$^{**}$ &  \\
##   &  &  &  &  & (0.062) & (0.075) &  \\
##   & & & & & & & \\
##  Constant & 2.358$^{***}$ & 2.448$^{***}$ & 1.907$^{***}$ & 3.000$^{***}$ & 3.029$^{***}$ & 2.914$^{*
##   & (0.058) & (0.075) & (0.089) & (0.471) & (0.135) & (0.164) & (0.071) \\
##   & & & & & & & \\
## \hline \\[-1.8ex]
## Observations & 117 & 117 & 116 & 117 & 117 & 117 & 116 \\
## Log Likelihood & $-$105.533 & $-$120.388 & $-$138.935 & $-$76.061 & $-$101.752 & $-$123.913 & $-$113
## Akaike Inf. Crit. & 217.066 & 246.776 & 283.869 & 162.121 & 211.504 & 255.826 & 232.417 \\
## Bayesian Inf. Crit. & 225.352 & 255.063 & 292.130 & 175.932 & 222.553 & 266.875 & 240.678 \\
```

```
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{7}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

As the anova tables show, three of the models for rubrics suggest that using the Rater Fixed Effect could be beneficial to the models. Also, one of the models suggests using the Sex fixed effect, which at least for now will be left out since it's only on one model. The rater fixed effect would make a lot of sense to be having some kind of influence on the rating as it is presented on the distribution of ratings by grader figure. If a grader is inclined to assign lower ratings compared to the others, it would be expected that the model can capture that behavior. The next step consists of using the Rater Fixed Effect model to construct new ICCs and compare to the previous. As the output shows, the same three models for the rubrics that suggested including the Rater Fixed Effect, which are **TxtOrg**, **InterpRes**, and **SelMethd** have a significant coefficient for it.

```
icc_df2 <- rep(NA, 7)
j = 1
for(i in c("RsrchQ", "CritDes",
           "InitEDA", "SelMeth",
           "InterpRes", "VisOrg",
           "TxtOrg")){
  form <- formula(paste0(i, "~ 1 + Rater +  (1 | Artifact)"))
  rnd_int <- lmer(form, data = ratings)
  # rnd_int <- fitLMER.fnc(rnd_int,
  #                        method = "BIC")
  rnd_int <- summary(rnd_int)
  icc_df2[j] <- rnd_int$varcor$Artifact/(rnd_int$varcor$Artifact + rnd_int$sigma^2)
  j = j+1
}

icc_df2 <- data.frame(Artifact = c("RsrchQ", "CritDes",
                                   "InitEDA", "SelMeth",
                                   "InterpRes", "VisOrg",
                                   "TxtOrg"),
                      ICC = icc_df2)
```

The ICCs for the model with Rater fixed effects have some effect on the original ICCs with some increases and decreases in them but no big changes on their magnitudes. This could suggest that including the Rater fixed effect is just helping get a better estimation for the ICC but the original models already did a good job.

Table 8: ICC for the models with variable selection

Artifact	ICC
RsrchQ	0.1967639
CritDes	0.6530379
InitEDA	0.7271755
SelMeth	0.5045807
InterpRes	0.1979727
VisOrg	0.6368757
TxtOrg	0.1646232

```
tall <- read.csv("/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project 2/tall.csv")
```

To be able to try if a Rubric Fixed Effect could be useful, a different structured data set with the same values was used. This way the Rubric is a categorical variable that can be added to the model. As Rater Fixed Effect was already part of the model, the Rubric was added to it to test how they work together. First, a base model with just the random effect for rubric was defined. Then, the model with Rubric and Rater Fixed Effects was fitted, having just the original random effects for Rubric. Finally, a model for each variable was tested to see if they could be used in the Random Effect. The intended method was a stepwise selection, but since it didn't work a model with a Random Effect for each variable was used for the **Rater**, **Repeated**, and **Semester** variables.

The results show that adding additional random effects makes the BIC greater, which means that they do not add explanatory power to the model with no random effects besides Rubric. This model only needs to be tested if it could use an interaction between the Rater and Rubric Fixed Effects.

```
#this is like an interaction between rubric and artifact for the RE. Zero means no intercept.
lmer.0 <- lmer(Rating ~ (0+Rubric|Artifact), data = tall)
ss <- getME(lmer.0,c("theta","fixef"))
lmer.0_u<- update(lmer.0,start=ss,
                  control=lmerControl(optimizer="nloptwrap",
                                      optCtrl=list(maxfun=2e5)))
lmer.1 <- lmer(Rating ~ Rater + Rubric +(1 + Rubric|Artifact),
               data = tall)
```

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

```
ss <- getME(lmer.1,c("theta","fixef"))
lmer.1<- update(lmer.1,start=ss,
                  control=lmerControl(optimizer="bobyqa",
                                      optCtrl=list(maxfun=2e5)))
```

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

```
lmer.2 <- lmer(Rating ~ Rater + Rubric +(1 + Rater+ Rubric|Artifact),
               data = tall)
```

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

```
ss <- getME(lmer.2,c("theta","fixef"))
lmer.2<- update(lmer.2,start=ss,
                  control=lmerControl(optimizer="Nelder_Mead",
                                      optCtrl=list(maxfun=2e5)))
```

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

```
lmer.3 <- lmer(Rating ~ Rater + Rubric +(1 + Repeated + Rubric|Artifact),
               data = tall)
lmer.4 <- lmer(Rating ~ Rater + Rubric+ Semester+(1 +  Rubric|Artifact),
               data = tall)
lmer.5 <- lmer(Rating ~ Rater + Rubric +Semester+(1 + Semester + Rubric|Artifact),
               data = tall)
```

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

```
anova (lmer.0_u,lmer.1, lmer.2, lmer.3, lmer.4, lmer.5)
```

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

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

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

## Data: tall
## Models:
## lmer.0_u: Rating ~ (0 + Rubric | Artifact)
## lmer.1: Rating ~ Rater + Rubric + (1 + Rubric | Artifact)
## lmer.4: Rating ~ Rater + Rubric + Semester + (1 + Rubric | Artifact)
## lmer.2: Rating ~ Rater + Rubric + (1 + Rater + Rubric | Artifact)
## lmer.3: Rating ~ Rater + Rubric + (1 + Repeated + Rubric | Artifact)
## lmer.5: Rating ~ Rater + Rubric + Semester + (1 + Semester + Rubric | Artifact)
##          npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## lmer.0_u   30 1537.2 1678.3 -738.58   1477.2
## lmer.1     37 1478.7 1652.8 -702.34   1404.7 72.4771  7  4.659e-13 ***
## lmer.4     38 1476.2 1655.0 -700.09   1400.2  4.4935  1  0.0340230 *
## lmer.2     45 1468.8 1680.6 -689.41   1378.8 21.3578  7  0.0032751 **
## lmer.3     45 1487.6 1699.4 -698.81   1397.6  0.0000  0
## lmer.5     46 1476.2 1692.7 -692.11   1384.2 13.3988  1  0.0002518 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, to see if interactions could help the model, an interaction between Rubric and Rater was added. This can potentially help to obtain different slopes for each Rubric and get better estimates.

The anova table that compares the model with no interaction and the model with interaction suggests that the second is a better fit for the data, which means that the selected model just uses the Rubric and Rater variables both as fixed and random effects.

```
lmer.1.1 <- lmer(Rating ~ Rater + Rubric +Rubric*Rater +(1 + Rubric|Artifact),
               data = tall)
lmer1.2 <- lmer(Rating ~ Rater + Rubric +Rubric*Rater +(1 + Rubric|Artifact),
               data = tall)
```

```
ss <- getME(lmer1.2,c("theta","fixef"))
lmer1.2.1<- update(lmer1.2,start=ss,
                   control=lmerControl(optimizer="Nelder_Mead",
                                       optCtrl=list(maxfun=2e5)))
```

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

```
anova(lmer.1, lmer1.2)
```

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

## Data: tall
## Models:
## lmer.1: Rating ~ Rater + Rubric + (1 + Rubric | Artifact)
## lmer1.2: Rating ~ Rater + Rubric + Rubric * Rater + (1 + Rubric | Artifact)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.1    37 1478.7 1652.8 -702.34   1404.7
## lmer1.2   43 1469.8 1672.1 -691.90   1383.8 20.881  6   0.001927 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final model is a random intercept and random slope with group level predictors. It looks like Rater and Rubric are the only factors that have strong influence in the ratings but also the combination of both. Since all artifacts have the same rubrics, it would be expected that they do not play a important role in the model, but when taken into account with rater, they become important if the raters are more inclined to assign certain ratings to rubrics, as shown in Q1. The interaction between rater and rubric shows that these holds true for the selected model, where most of the interactions are significant (for the purpose of this paper, a coefficient is considered significant if the absolute value of their t-value is greater than two). It is also interesting to note that adding the interaction takes significance off rater, meaning that the rater alone does not have influence on the rating but the patterns that each rater show for the rubrics do.

```
summary(lmer1.2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + Rubric + Rubric * Rater + (1 + Rubric | Artifact)
##    Data: tall
##
## REML criterion at convergence: 1437.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9956 -0.5169 -0.0412  0.4940  3.6064
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact (Intercept)     0.5337   0.7306
##           RubricInitEDA   0.4716   0.6867   -0.65
##           RubricInterpRes 0.4570   0.6760   -0.84  0.87
##           RubricRsrchQ    0.3149   0.5612   -0.84  0.63  0.90
##           RubricSelMeth   0.4103   0.6405   -0.92  0.76  0.93  0.81
##           RubricTxtOrg    0.4823   0.6945   -0.75  0.76  0.86  0.78  0.84
```

```
##           RubricVisOrg    0.4918   0.7013   -0.73  0.83  0.86  0.75  0.76  0.87
##  Residual                 0.1883   0.4339
## Number of obs: 817, groups:  Artifact, 91
##
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)             1.68009    0.16027  10.483
## Rater                   0.11390    0.06697   1.701
## RubricInitEDA           0.84696    0.19650   4.310
## RubricInterpRes         1.33884    0.19403   6.900
## RubricRsrchQ            0.83371    0.18213   4.577
## RubricSelMeth           0.55987    0.18805   2.977
## RubricTxtOrg            1.18540    0.19727   6.009
## RubricVisOrg            0.87126    0.19869   4.385
## Rater:RubricInitEDA    -0.15279    0.08633  -1.770
## Rater:RubricInterpRes  -0.37749    0.08519  -4.431
## Rater:RubricRsrchQ     -0.18934    0.08089  -2.341
## Rater:RubricSelMeth    -0.20011    0.08264  -2.422
## Rater:RubricTxtOrg     -0.24904    0.08657  -2.877
## Rater:RubricVisOrg     -0.17320    0.08700  -1.991

##
## Correlation matrix not shown by default, as p = 14 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

4

Finally, using a regular linear model helps understand the influence of the explanatory variables in the outcome of the ratings. It is worth noting that all of the variables except for **Repeated** have significant coefficients, which means that at some point all of them can help explain the variation of **Ratings**. Using this model can help for further MLM analyses with the current data. Also understand the differences between the variables is useful to potentially help the MLM model point in the correct direction as some of the commands depend on Convex Optimization and fail from time to time. Knowing that most of the variables could help the model suggests that the study could take a step back to analyze if any of them could be added to the MLM model's fixed effects or why not.

```
tall$Rater <- as.factor(tall$Rater)
tall$Repeated <- as.factor(tall$Repeated)
tall$Semester <- as.factor(tall$Semester)
tall$Sex <- as.factor(tall$Sex)
tall$Rubric <- as.factor(tall$Rubric)
lm_ratings <- lm(Rating~.--1,
                 data = tall %>%
                   dplyr::select(-Artifact, -X))
summary(lm_ratings)
```

```
##
## Call:
## lm(formula = Rating ~ . - -1, data = tall %>% dplyr::select(-Artifact,
##     -X))
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78582 -0.46075 -0.07607  0.47996  2.07886
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.74878    0.25818  10.647  < 2e-16 ***
## Rater2           0.07645    0.05581   1.370 0.171125
## Rater3          -0.19567    0.05614  -3.485 0.000518 ***
## Repeated1       -0.07213    0.04856  -1.485 0.137851
## SemesterS19     -0.13751    0.05248  -2.620 0.008956 **
## SexF            -0.76658    0.25148  -3.048 0.002377 **
## SexM            -0.79784    0.25089  -3.180 0.001529 **
## RubricInitEDA    0.56478    0.08519   6.630 6.17e-11 ***
## RubricInterpRes  0.61606    0.08519   7.232 1.11e-12 ***
## RubricRsrchQ     0.47931    0.08519   5.626 2.54e-08 ***
## RubricSelMeth    0.19726    0.08519   2.316 0.020836 *
## RubricTxtOrg     0.72717    0.08519   8.536  < 2e-16 ***
## RubricVisOrg     0.54363    0.08537   6.368 3.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6502 on 804 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1446
## F-statistic:  12.5 on 12 and 804 DF,  p-value: < 2.2e-16
```

```
lm_ratings2 <- stepAIC(lm_ratings)
```

```
## Start:  AIC=-690.6
## Rating ~ (Rater + Repeated + Semester + Sex + Rubric) - -1
##
##            Df Sum of Sq    RSS     AIC
## <none>                  339.86 -690.60
## - Repeated  1     0.933 340.79 -690.36
## - Semester  1     2.902 342.76 -685.65
## - Sex       2     4.339 344.20 -684.23
## - Rater     2    10.561 350.42 -669.60
## - Rubric    6    45.830 385.69 -599.25
```

```
summary(lm_ratings2)
```

```
##
## Call:
## lm(formula = Rating ~ (Rater + Repeated + Semester + Sex + Rubric) -
##     -1, data = tall %>% dplyr::select(-Artifact, -X))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78582 -0.46075 -0.07607  0.47996  2.07886
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      2.74878    0.25818  10.647  < 2e-16 ***
## Rater2           0.07645    0.05581   1.370 0.171125
## Rater3          -0.19567    0.05614  -3.485 0.000518 ***
## Repeated1       -0.07213    0.04856  -1.485 0.137851
## SemesterS19     -0.13751    0.05248  -2.620 0.008956 **
## SexF            -0.76658    0.25148  -3.048 0.002377 **
## SexM            -0.79784    0.25089  -3.180 0.001529 **
## RubricInitEDA    0.56478    0.08519   6.630 6.17e-11 ***
## RubricInterpRes  0.61606    0.08519   7.232 1.11e-12 ***
## RubricRsrchQ     0.47931    0.08519   5.626 2.54e-08 ***
## RubricSelMeth    0.19726    0.08519   2.316 0.020836 *
## RubricTxtOrg     0.72717    0.08519   8.536  < 2e-16 ***
## RubricVisOrg     0.54363    0.08537   6.368 3.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6502 on 804 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.1572, Adjusted R-squared:  0.1446
## F-statistic:  12.5 on 12 and 804 DF,  p-value: < 2.2e-16
```

```
lm_ratings3 <- lm(Rating~Rater + Repeated + Semester + Sex + Rubric + Rubric*Rater-1,
                  data = tall %>%
                    dplyr::select(-Artifact, -X))
summary(lm_ratings3)
```

```
##
## Call:
## lm(formula = Rating ~ Rater + Repeated + Semester + Sex + Rubric +
##     Rubric * Rater - 1, data = tall %>% dplyr::select(-Artifact,
##     -X))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8169 -0.4361 -0.0745  0.4699  1.9297
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## Rater1                    2.43548    0.26713   9.117  < 2e-16 ***
## Rater2                    2.97384    0.26764  11.112  < 2e-16 ***
## Rater3                    2.72161    0.26129  10.416  < 2e-16 ***
## Repeated1                -0.07224    0.04807  -1.503 0.133261
## SemesterS19              -0.13681    0.05195  -2.633 0.008618 **
## SexF                     -0.76671    0.24893  -3.080 0.002142 **
## SexM                     -0.79803    0.24835  -3.213 0.001365 **
## RubricInitEDA             0.82051    0.14574   5.630 2.50e-08 ***
## RubricInterpRes           1.12821    0.14574   7.741 3.01e-14 ***
## RubricRsrchQ              0.84615    0.14574   5.806 9.27e-09 ***
## RubricSelMeth             0.53846    0.14574   3.695 0.000235 ***
## RubricTxtOrg              1.17949    0.14574   8.093 2.19e-15 ***
## RubricVisOrg              0.80707    0.14670   5.502 5.09e-08 ***
## Rater2:RubricInitEDA     -0.38642    0.20679  -1.869 0.062039 .
## Rater3:RubricInitEDA     -0.38462    0.20611  -1.866 0.062401 .
## Rater2:RubricInterpRes   -0.66847    0.20679  -3.233 0.001277 **
```

```
## Rater3:RubricInterpRes -0.87179    0.20611  -4.230 2.61e-05 ***
## Rater2:RubricRsrchQ    -0.61719    0.20679  -2.985 0.002927 **
## Rater3:RubricRsrchQ    -0.48718    0.20611  -2.364 0.018334 *
## Rater2:RubricSelMeth   -0.54026    0.20679  -2.613 0.009155 **
## Rater3:RubricSelMeth   -0.48718    0.20611  -2.364 0.018334 *
## Rater2:RubricTxtOrg    -0.71975    0.20679  -3.481 0.000528 ***
## Rater3:RubricTxtOrg    -0.64103    0.20611  -3.110 0.001937 **
## Rater2:RubricVisOrg    -0.29606    0.20747  -1.427 0.153971
## Rater3:RubricVisOrg    -0.49938    0.20679  -2.415 0.015964 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6436 on 792 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9316, Adjusted R-squared:  0.9294
## F-statistic: 431.3 on 25 and 792 DF,  p-value: < 2.2e-16
```

```
stargazer::stargazer(summary(lm_ratings3)$coef, type = "latex", summary = F, digits = 3)
```

The models from part 2 gave access to ICC. This gives a standard way of seeing the agreement between graders, but it's not enough to see if they just grade exactly the same or in the same direction.

There are situations when there is variation between one group to the next, so in this cases it would be understandable to use a fixed effect (in the context of lm). In this data one can group by artifact, rubric and grader. If the raters could be inconsistent, grouping RE by them could not be showing really random differences between them (systematic differences). Rubric is not a good RE because of systematic differences: some people can actually be better at some of them.

Try a model with rater as grouping effect. Try a model with groups by rater and groups by artifact.

It is only expected to use artifact as grouping variable.