# An Analysis and Evaluation of Dietrich College's New General Education Program

Kevin Yang | kevinyan@andrew.cmu.edu | Department of Statistics & Data Science, Carnegie Mellon University

## Abstract

This study aims to answer several questions from Carnegie Mellon University regarding the success of the implementation of a new "General Education"/ "Gen Ed" program for undergraduates based on the quality of how papers are rated. Data for this study was obtained through 91 randomly selected papers, also called artifacts in this study, rated by 3 Raters on 7 different criteria, also called Rubrics in this study, on a scale from 1 to 4. Analysis of this data was conducted through Exploratory Data Analysis (EDA) and linear mixed-effects models to find similarities and differences between the Raters' and Rubrics' rating distributions, and to find any factors that influence the rating of a paper for each of the 7 Rubrics. Consistent rating patterns can be found when grouping the ratings by Rater, but inconsistent rating patterns were when the ratings by Rubric. Additionally, for some of the Rubrics, 3 of them to be exact, no external factors were found to influence the ratings of a paper, but for the other 4 Rubrics, various factors such as the Rater, and Semester were found necessary in predicting the rating of an artifact. Overall, the results from this study show that the Gen Ed program has a few flaws in the rating system. In particular, the rating meanings can be interpreted several different ways depending on the Rubric or the Rater which causes inconsistent and biased ratings.

## Introduction

The Dietrich College in Carnegie Mellon University is currently implementing a new "General Education" program for undergraduates to give each student a baseline knowledge of certain subjects. In order to measure the success of this program, the college will be rating papers written by students on several key criteria/Rubrics. As such, the college has been curious with how ratings are currently given out to students when they write artifacts, are ratings being given out fairly and are unbiased towards students? Do the ratings reflect the content written in an artifact, or does an external factor such as the Rater rating the artifact have a larger influence on the rating? In an ideal setting, ratings should be distributed normally and not be dependent on anything other than the contents of the artifact to ensure students receive the rating they deserve. For this study, past artifacts and their ratings will be analyzed to answer four questions:

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

## Data

The data for this study directly comes from Brian Junker (2021) at Carnegie Mellon University where 91 papers/artifacts were sampled from a freshman statistics class. Three Raters/graders from different departments were asked to rate these papers on 7 different criteria/Rubrics on a scale of 1 to 4 with 4 being the best. Below are two tables explaining the Rubrics and the rating scale.

Rubric Names and Descriptions

| Short Name | Full Name | Description |
|---|---|---|
| RsrchQ | Research Question | Given a scenario, the student generates, critiques or evaluates a relevant empirical research question. |
| CritDes | Critique Design | Given an empirical research question, the student critiques or evaluates to what extent a study design convincingly answers that question. |
| InitEDA | Initial EDA | Given a data set, the student appropriately describes the data and provides initial Exploratory Data Analysis. |
| SelMeth | Select Methods | 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.). |

Rating Meanings

| 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 |

In addition to the ratings, various other external factors were included for each artifact, detailed below:

Non-Rubric Variable Meanings

| Variable Name | Values | Description |
|---|---|---|
| (X) | 1,2,3... | Row number in dataset |
| Rater | 1,2,3 | Which Rater gave the rating |
| (Sample) | 1,2,3... | Sample number |
| (Overlap) | 1,2,3…,13 | Identifier for artifact seen by all three Raters |
| Semester | Fall or Spring | Which semester the artifact was written |
| Sex | M or F | Gender of the student |
| Artifact | (Text labels) | Unique identifier for each artifact |
| Repeated | 0 or 1 | 1=An overlap artifact |

In the table above, any variable or value contained within parentheses are not meaningful and won't be used in the study. Of the remaining variables, the data was presented in two tables, the first table is called "rating" where each Rater and artifact has its own row and each Rubric has its own column with the given rating. The second table is called "tall" and only has one column specifying the Rubric and another column specifying the rating, meaning each artifact has multiple rows for each Rater and Rubric. On top of that, both tables will be subsetted with the artifacts that were viewed by all three Raters either by if Repeated has a value of 1 or if the Overlap column has a value present. In total, there will be four datasets used for this study. Summary statistics for the "rating" dataset are shown below:

Summary Statistics for Numeric Variables in the “Rating” Dataset

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | SD |
|---|---|---|---|---|---|---|---|
| Rater | 1 | 1 | 2 | 2.01 | 3.0 | 3 | 0.82 |
| 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 |

Summary Statistics for Semester and Sex Variables in “Rating” Dataset

| | Fall | Spring |
|---|---|---|
| Freq | 83 | 34 |

| | – | F | M |
|---|---|---|---|
| Freq | 1 | 64 | 52 |

## Methods

### Method 1 - Rating Distributions

To begin, the first question regarding the distribution of ratings by Rubric and by Rater only requires some simple Exploratory Data Analysis. Bar charts were created to visualize the distribution of ratings based on each Rubric and each Rater for both the overlapping artifacts and the full dataset. To give more details about these bar charts, tables of means and standard deviations for each Rubric and Rater were made on both datasets to compare the distributions for each Rubric and each Rater.

### Method 2 - Rater Agreement

In the second question about if the Raters agree with each other, something called the intraclass correlation (ICC) was calculated to compare the Raters on both the overlapping and full dataset for all the Rubrics. The ICC calculates how often all three Raters give the same rating to a particular artifact on a particular Rubric in the form of a percentage. Next, the Percent Exact Agreement was calculated for each pair of Raters (1 & 2, 1 & 3, 2 & 3) on the overlapping dataset by comparing the frequencies of each rating by each Rater on each Rubric and counting how many of the 13 artifacts received the same rating.

This count was then converted to a percentage to see what proportion of the 13 artifacts received the same rating by two Raters on a particular Rubric. All of these results were combined into a comprehensive table with each Rubric as a row and five columns containing the ICCs for the overlapping dataset, the ICCs for the full dataset, and the percent exact agreements between Raters 1 and 2, 1 and 3, and 2 and 3.

### Method 3 - Factor Influence on Ratings

When considering all of the various factors in this study such as the Rater, Semester, Sex, Overlapping artifacts, and each Rubric, finding a model to predict an artifact's rating can be a very tedious task considering every possible combination of factors. For this question, we will be using three types of "terms" in our model: fixed effects, random effects, and interaction terms. Fixed effects are predictors that do not consider any grouping in the data and looks at the dataset as a whole. Random effects are predictors that do consider grouping in the dataset, in this case, each artifact. Interaction terms are predictors that explain any relationship between two other predictors that could influence the rating of an artifact that neither fixed nor random effects can detect.

Completing this task will be done in multiple steps, first, every combination of fixed effects (Rater, Semester, Sex, and Repeated) will be fitted on a random intercept model to see which fixed effects had the most influence on the rating on both the overlapping and the full dataset across all Rubrics at once. Next, the fixed effects models were fitted for each Rubric individually and the most significant models according to ANOVA will be used as the model to predict the rating. This will be done twice, once on the overlapping dataset and once on the full dataset. If any of the models have terms other than a random intercept, significance tests will be run to see if they are significant and if interaction terms are necessary if there is more than one fixed effect in the chosen model. Finally, one final model will be fitted containing every fixed effect and interaction term possible just to see if there are any interesting interactions between the factors that weren't detected by the previous model fitting methods.

### Method 4 - Additional Analysis

For the final part of the study, a couple other EDA plots were made to see if there was anything else that could be deemed interesting for this study. In this case, two boxplots were created to compare the distribution of total ratings based on genders and based on semesters across all of the Rubrics. For this part, because of the use of the sum of the artifacts' ratings, artifacts that contained an "NA" rating in any Rubric were removed from the dataset.

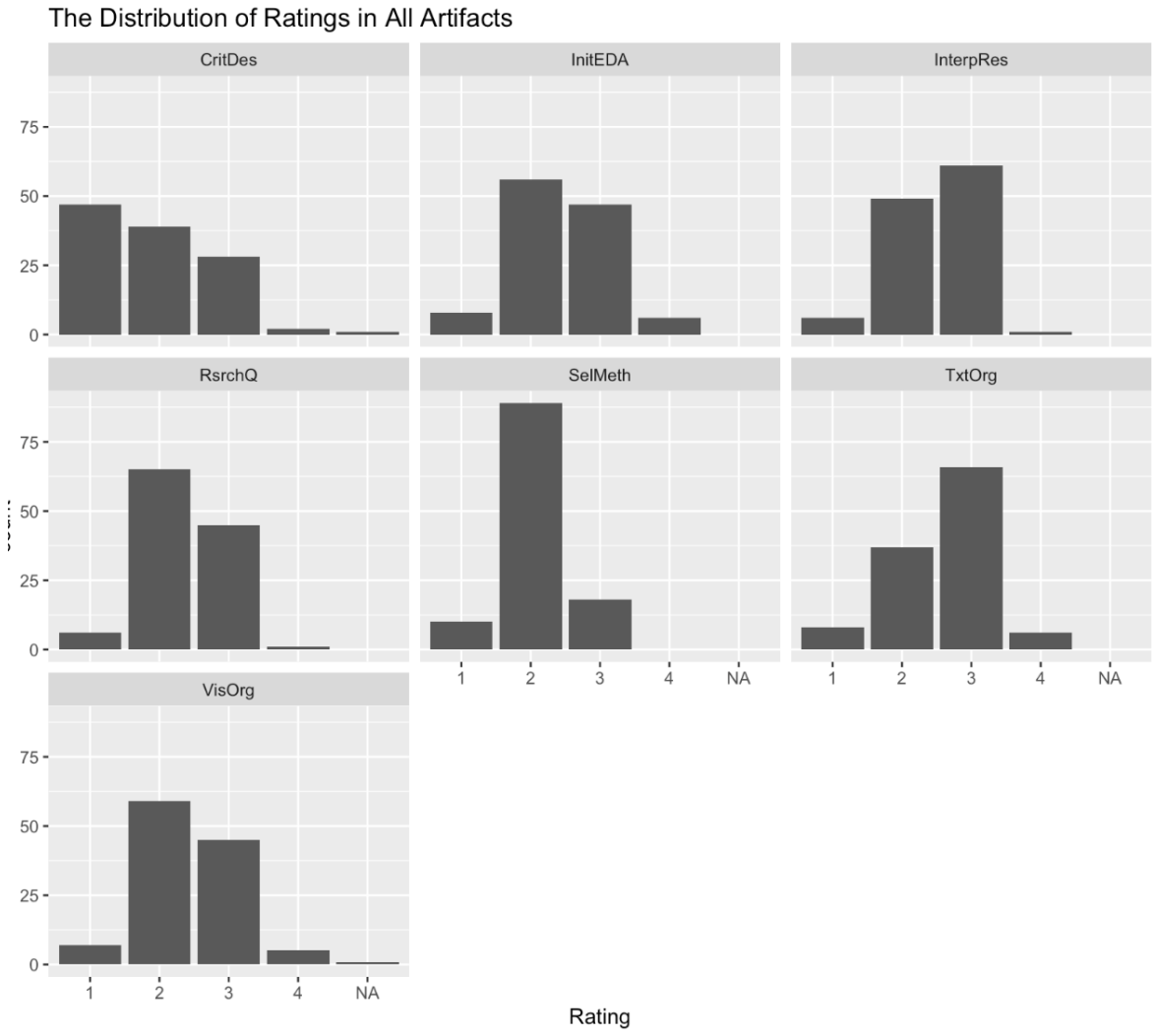
## Results

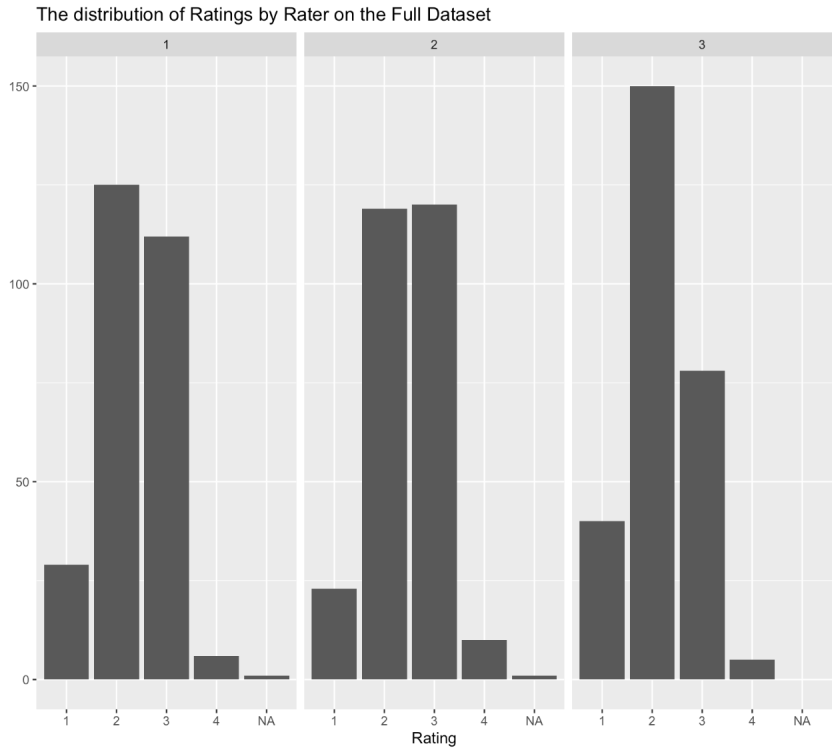
### Results 1 - Raters rate very similarly to each other while Rubrics have inconsistent distributions for the ratings

Looking at the bar plots in Appendix A and below, it appears that when the ratings are grouped by Rubric, their distributions are very different from each other. The Initial EDA, Research Question, and Visual Organization Rubrics make up a group of very similar distributions where there is a high frequency of 2's, followed by a slightly smaller frequency of 3's and very few 1's and 4's forming a normal distribution with a nice spread. However, the remaining four Rubrics do not follow this pattern. The Critique Design Rubric for example has a high frequency of low ratings, mostly 1's and 2's, the Interpret

Results and Text Organization Rubrics have more 3's than 2's, and the Select Methods Rubric has an extremely high frequency of 2's.

On a positive note, in the plots showing the ratings grouped by Rater, it appears that all three Raters gave a similar distribution of ratings on all of the artifacts with a high frequency of 2's and 3's, a smaller frequency of 1's and a miniscule frequency of 4's. Below are the barplots for the full dataset, the barplots for the overlapping dataset can be found in the Appendix A.






More details about these distributions can be seen in the mean and standard deviation tables in Appendix B. To summarize, it appears that Rater 3 tends to give lower ratings on average as seen in the first two tables and that Rater 1 has a smaller spread of ratings in most of the Rubrics. In addition, the average rating for the Critique Design Rubric is noticeably lower than the other Rubrics and has a larger spread of ratings disregarding the Raters.

**Results 2 - Raters don't always agree well enough on what rating an artifact should receive**

In the table below and in Appendix D, a table of ICCs and Percent Exact Agreement can be found to answer the second question. Each number in this table represents a percentage, so 0.19 means 19% agreement. The process for calculating the Percent Exact Agreement can be found in Appendix C.

Table 5: ICC and Percent Agreement for each Rubric and Pair of Raters

| Rubric | ICC for Overlaps | ICC for Full | Rater 1 & 2 | Rater 1 & 3 | Rater 2 & 3 |
|---|---|---|---|---|---|
| RsrchQ | 0.19 | 0.21 | 0.38 | 0.77 | 0.54 |
| CritDes | 0.57 | 0.67 | 0.54 | 0.62 | 0.69 |
| InitEDA | 0.49 | 0.69 | 0.69 | 0.54 | 0.85 |
| SelMeth | 0.52 | 0.47 | 0.92 | 0.62 | 0.69 |
| InterpRes | 0.23 | 0.22 | 0.62 | 0.54 | 0.62 |
| VisOrg | 0.59 | 0.66 | 0.54 | 0.77 | 0.77 |
| TxtOrg | 0.14 | 0.19 | 0.69 | 0.62 | 0.54 |

Looking at the ICCs, it appears that all three Raters disagree on the scoring of the Research Question, Interpret Results, and Text Organization Rubrics, each having ICCs of around 20% on both the overlapping artifacts and the full dataset. For the remaining Rubrics, the ICCs don't go any higher than 70%. Between each pair of Raters, their Percent Exact Agreements are usually in the range of 50% and 80% for most of the Rubrics, the only main exceptions are between Raters 1 and 2 on the Research Question Rubric with an agreement of 38% and on the Select Methods Rubric with an agreement of 92%.

**Results 3 - The ratings for some Rubrics are dependent on the Rater and the Semester**

For the third question, when looking at only the artifacts in the Overlapping dataset across all of the Rubrics at once, the best model to predict an artifact's rating is a random intercept model. More information on how this model was chosen can be found in Appendices E and F, but in simplest terms, none of the fixed effects Rater, Semester, Sex, or Repeated were found to be significant in an overlapping artifact's ratings. A random intercept model for a Rubric consists of a single "intercept" term, usually the average rating for that Rubric along with two "error" terms that accounts for the randomness that can change an artifact's rating. The two error terms are for if the ratings are grouped by artifact or not. When expanding the dataset to include all artifacts, Rater suddenly becomes a significant fixed effect so it has some influence on the rating. Like before the process for choosing this model can be found in Appendices G, H and I.

Next, when creating models for each Rubric separately, the overlapping artifacts show that the best model for each Rubric is a random intercept model similar to how when all of the Rubrics were grouped together. However, when creating the models for all of the artifacts, only three of the seven Rubrics resulted in a random intercept model; these were Initial EDA, Research Question, and Text Organization. This means that their ratings can be determined by an average with an error. For the remaining four Rubrics, other fixed effects were found to be significant in predicting the ratings. For the Critique Design, Interpret Results, and Visual Organization Rubrics, the Rater was a significant fixed effect, and for the Select Methods Rubric, the Rater and Semester were both significant fixed effects with an insignificant interaction term. A table of all of the coefficients can be found below and in Appendix M:

Coefficients of the Final Models for each Rubric

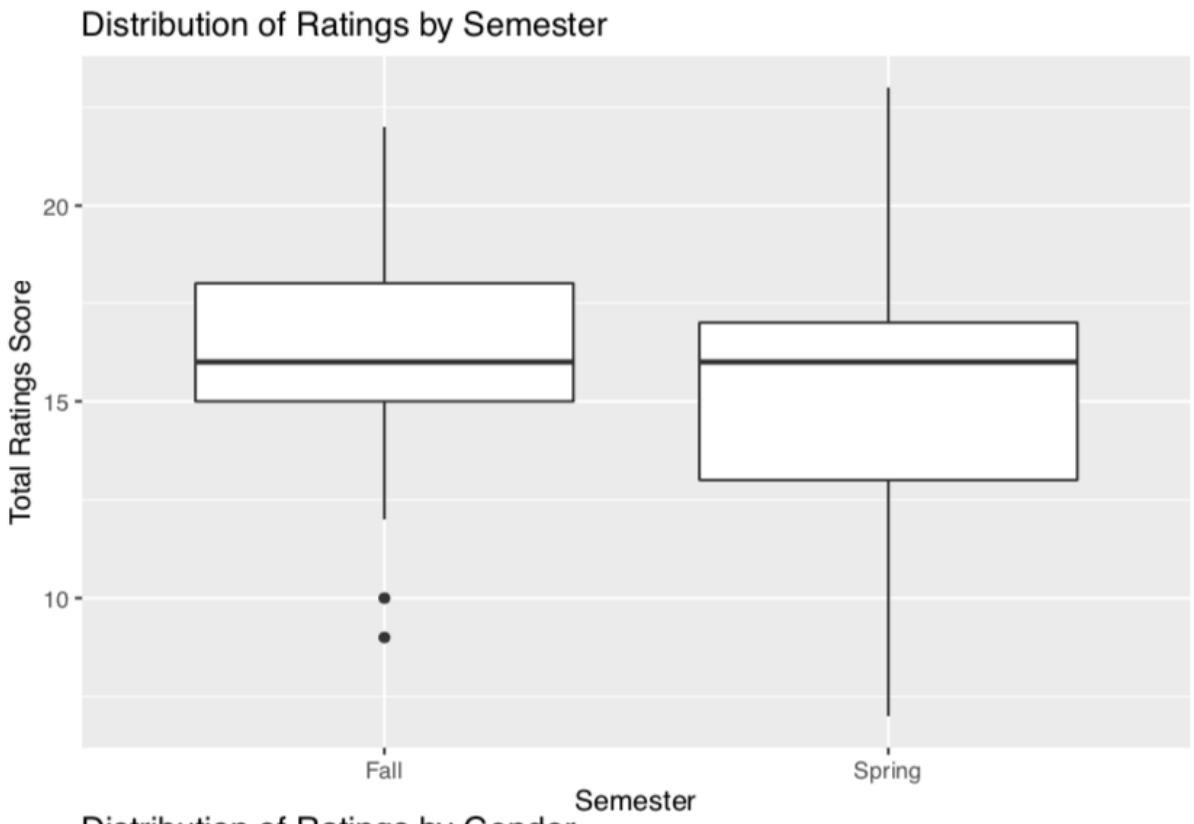
| | RsrchQ | CritDes | InitEDA | SelMeth | InterpRes | VisOrg | TxtOrg |
|---|---|---|---|---|---|---|---|
| Intercept | 2.35 | 0 | 2.44 | 0 | 0 | 0 | 2.59 |
| Rater 1 | 0 | 1.69 | 0 | 2.25 | 2.70 | 2.38 | 0 |
| Rater 2 | 0 | 2.11 | 0 | 2.23 | 2.59 | 2.65 | 0 |
| Rater 3 | 0 | 1.89 | 0 | 2.03 | 2.14 | 2.28 | 0 |
| SemesterS19 | 0 | 0 | 0 | -0.36 | 0 | 0 | 0 |
| Variance by Artifact ($\tau^2$) | 0.07 | 0.43 | 0.37 | 0.09 | 0.06 | 0.29 | 0.09 |
| Variance of Residuals ($\sigma^2$) | 0.28 | 0.24 | 0.17 | 0.11 | 0.25 | 0.15 | 0.40 |

Each row in this table represents a term in the linear model and each column represents the rating from a particular Rubric. With the exception of the Select Methods Rubric, each non-zero value in this table represents the average rating for an artifact for a particular Rubric. A non-zero value in the intercept row means that the average applies to all artifacts regardless of Rater. Non-zero values in the Rater 1, Rater 2, and Rater 3 rows represent the average rating for an artifact for a particular Rubric for that specific Rater. For the Select Methods Rubric, artifacts rated in the Spring Semester were rated 0.36 points lower than in the Fall so Rater averages for the Spring Semester are 0.36 points less than what is shown on the table. The two rows of variances at the bottom of the table indicate the average difference in rating for each artifact on a particular Rubric. Smaller variances means that all of the ratings are very similar and close to each other in value while larger variances mean that the ratings are farther apart on average. Variance by Artifact covers the rating differences after grouping ratings by artifacts, and Variances of Residuals covers the rating difference independently from a rating's artifact, both are necessary in predicting an artifact's rating.

For the last part of the third question where a large model was created to see if there were any other interesting interactions between the factors on the full dataset, 11 combinations of factors were found significant, all consisting of a Rater and some combination of the other factors. This part is purely exploratory and the tables are very large so the table containing the significant interaction terms can be found in Appendix N along with a general interpretation of the table values.

### Results 4 - The distribution of ratings by student gender are nearly identical while the ratings by semester are skewed a little bit

Some other interesting results from the data to answer the fourth question are that the median total rating across all of the Rubrics are about the same when grouping the artifacts by either semester or gender as seen by the boxplots. Differences in the rating distributions can mostly be found in the spread of the boxplot as there appears to be a wider spread of ratings in the spring semester than in the fall, and a slightly wider spread among females than among males. Another interesting observation from the boxplots is that the distribution of ratings by gender is symmetrical for both groups while the distributions by semester is skewed up for the fall and skewed down for the spring. Looking at the counts and averages for each grouping, the average total rating across all of the Rubrics are nearly the same with females having a slightly higher average by about 0.02 points. Between semesters there is a larger difference in the average total rating, with the fall semester average being noticeably higher than the spring, by about 0.87 points. The boxplots and tables can be seen below and in Appendix O.


Distribution of Ratings by Semester
Total Ratings Score
20
15
10
Fall
Spring
Semester

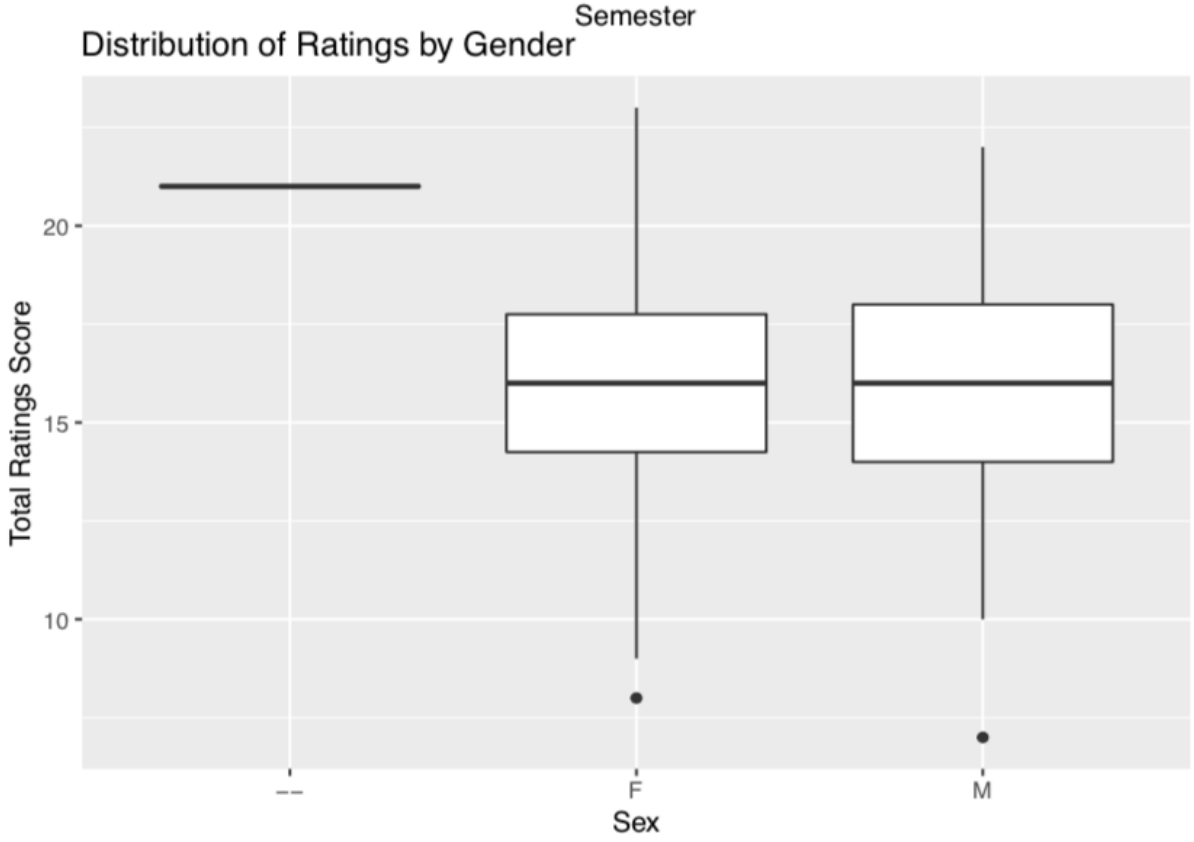

Distribution of Ratings by Gender
Total Ratings Score
20
15
10
--
F
M
Sex

Table 21: Average Rating of Artifacts by Gender

| Sex | genderrating | gendermean | count |
|---|---|---|---|
| – | 21 | 21.00000 | 1 |
| F | 1004 | 16.19355 | 62 |
| M | 841 | 16.17308 | 52 |

Table 22: Average Rating of Artifacts by Semester

| Semester | semesterrating | semestermean | count |
|---|---|---|---|
| Fall | 1351 | 16.47561 | 82 |
| Spring | 515 | 15.60606 | 33 |

## Discussion

The purpose of this study was to answer four questions posed by Dietrich College at Carnegie Mellon University regarding their new "General Education" program: whether the distribution of ratings is the same across each Rubric and Rater, if Raters tend to give the same scores for the same Rubric on the same artifact, how internal factors such as the contents of an artifact affect the rating compared to external factors such as the person rating the artifact, the student's sex or the semester the artifact was written, and if there are any other interesting conclusions that can be deduced from the data. The ratings in this study are being used to measure the success of the Gen Ed program and so it is important to look for patterns in the data to ensure that the rating process is consistent and unbiased.

One of the positives that came out of this study was the consistency of the Raters when it came to rating the papers. Their distribution of ratings were mostly consistent across all of the Rubrics independent of non-Rubric factors. This is a good start because it means that the Raters are rating fairly and aren't giving any students an unfair advantage or disadvantage. This can be argued further using the nearly identical boxplots from the last research question because it demonstrates that the Raters are unbiased towards gender when it comes to ratings. The only sign of bias from the ratings appears in the semester factor as there is a wider spread of ratings in the spring than in the fall. However, while the boxplot spreads might be a little concerning at first, it's entirely likely that there are confounding factors involved here because class sizes are significantly different from each other, as such it's likely harder to normally distribute ratings in a class of 30 compared to a class of 80. Ultimately, for the Gen Ed program to be successful, the Raters should be as unbiased as possible when rating artifacts and they're doing mostly a good job so far given the evidence provided.

On the other hand, one of the biggest negatives that came out of this study came from the point of view of the Rubrics. When the frame of reference for the data changed from the Rater to the Rubrics, much of the consistency disappeared and it was clear that more clarity is needed for how Rubrics are scored. Given the inconsistencies and differences of the bar plot distributions, the low ICCs between the Raters, and the presence of the Rater fixed effect in the models predicting rating, it is highly likely that the Raters are interpreting the rating meanings in relation to the Rubric descriptions. The source of the issue possibly lies in the descriptions of the Rating meanings such as "significant flaws", "competent", and

"outstanding" because these words are hard to quantify and are up to the Rater's opinion to determine what counts as a 1, 2, 3, or 4. While there is no simple way to eliminate these inconsistencies, one possible solution for this is to create rating meanings specific to each Rubric so Raters can have a bit more clarity and guidance on what is expected of the students, and how to separate "middle grades" like a 2, or a 3. I've provided a rough draft example of this below for the Select Methods Rubric.

Rubric Meanings for SelMeth (Can be modified)

| Rating | Description |
|---|---|
| 1 | No methods mentioned or irrelevant work provided |
| 2 | Student provides a broad method/concept to analyze data (e.g. Specifying regression without mentioning the type of regression) |
| 3 | Student provides one specific and detailed method to conduct data analysis |
| 4 | Student provides multiple related, specific, and detailed methods to conduct data analysis |

There are several ways that better analyses and further research can be conducted on the data. To begin, some of the artifacts had "NA" values for some of the Rubric ratings and were removed from the dataset during parts of the study such as for the boxplots in research question 4. There was no indication in the dataset as to why these artifacts received "NA" ratings and these reasons could be factors that influence the results of the study. In the future, it might be beneficial to either only use artifacts that have ratings in all of the Rubrics, or to add clarity to artifacts with missing or unusual ratings. Next, there wasn't a lot of data to work with as 91 artifacts for the full dataset doesn't seem to be a sufficient amount of data, especially since only 13 of the artifacts were reviewed by all three Raters. Including more artifacts in the dataset, with more artifacts that are rated by all of the Raters, and possibly including artifacts from past semesters will give more solid analyses and insights on how ratings are distributed and how Raters differ in rating the same artifacts. Lastly, given that the artifacts for this study were from a freshman statistics class, and that the Raters come from different departments in Dietrich College, it would be interesting to see how the ratings would change if all of the Raters came from the same department or were all from the Statistics department. It seems plausible that the interpretation differences speculated above could come from the Raters' various backgrounds like English, History, Economics, Information Systems, or the other departments in Dietrich College. Adding a "Rater department" column to the data or as a fixed or random effect could provide more insight into the distribution of ratings.

Overall, the rating system made for the "General Education" program needs several improvements right now, as there is a great amount of inconsistencies when it comes to how ratings are given out. Even though each Rater is giving out ratings fairly and in an unbiased manner, the distribution of ratings based on Rubric is often biased by Rater and how they interpret the Rubric meanings. In order to ensure that students fairly receive the rating they are entitled to, more clarity should be described on the Rubrics and rating meanings so that the Raters can have better guidance on how to rate a student's work.

## References

Junker, B. W. (2021). *Project 02 assignment sheet and data for 36-617: Applied Regression Analysis.* Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh PA. Accessed Nov 15, 2021 from https://canvas.cmu.edu/courses/25337/files/folder/Project02

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

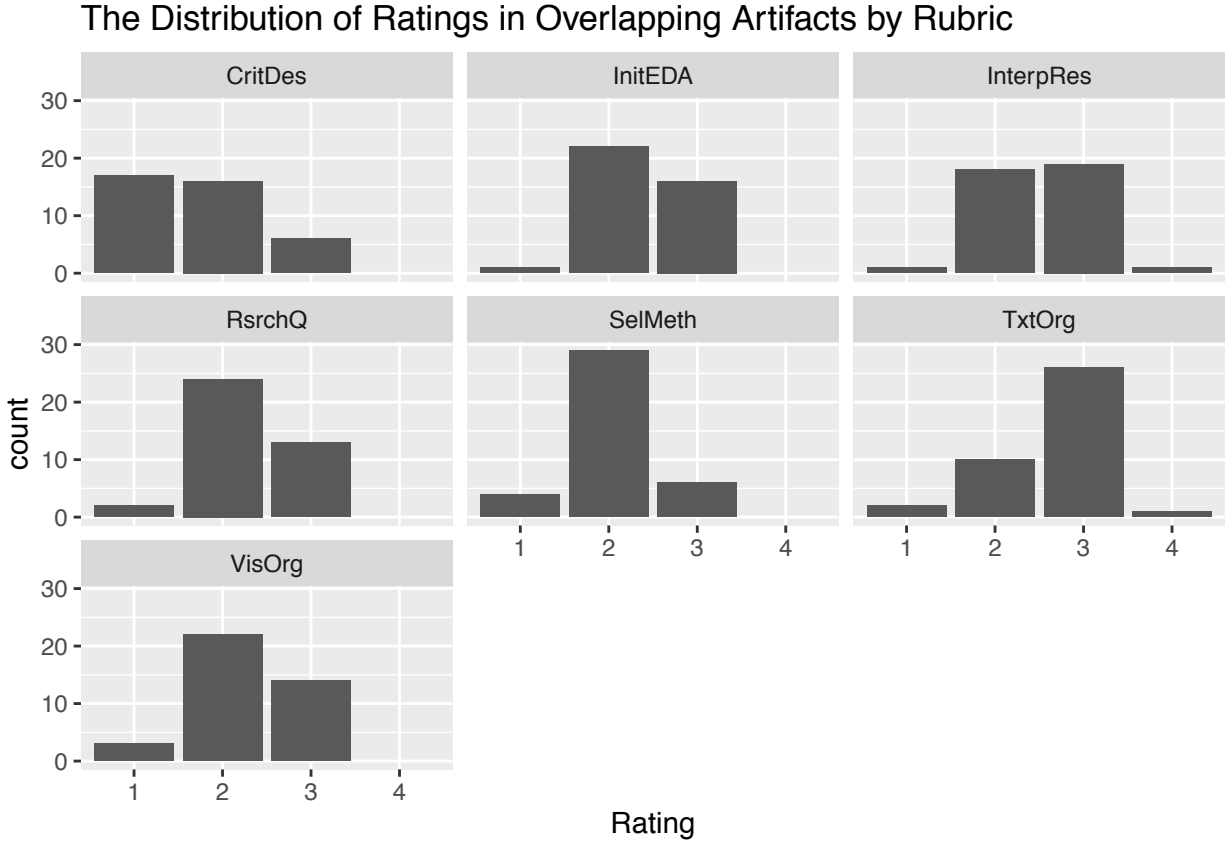
# Technical Appexdix

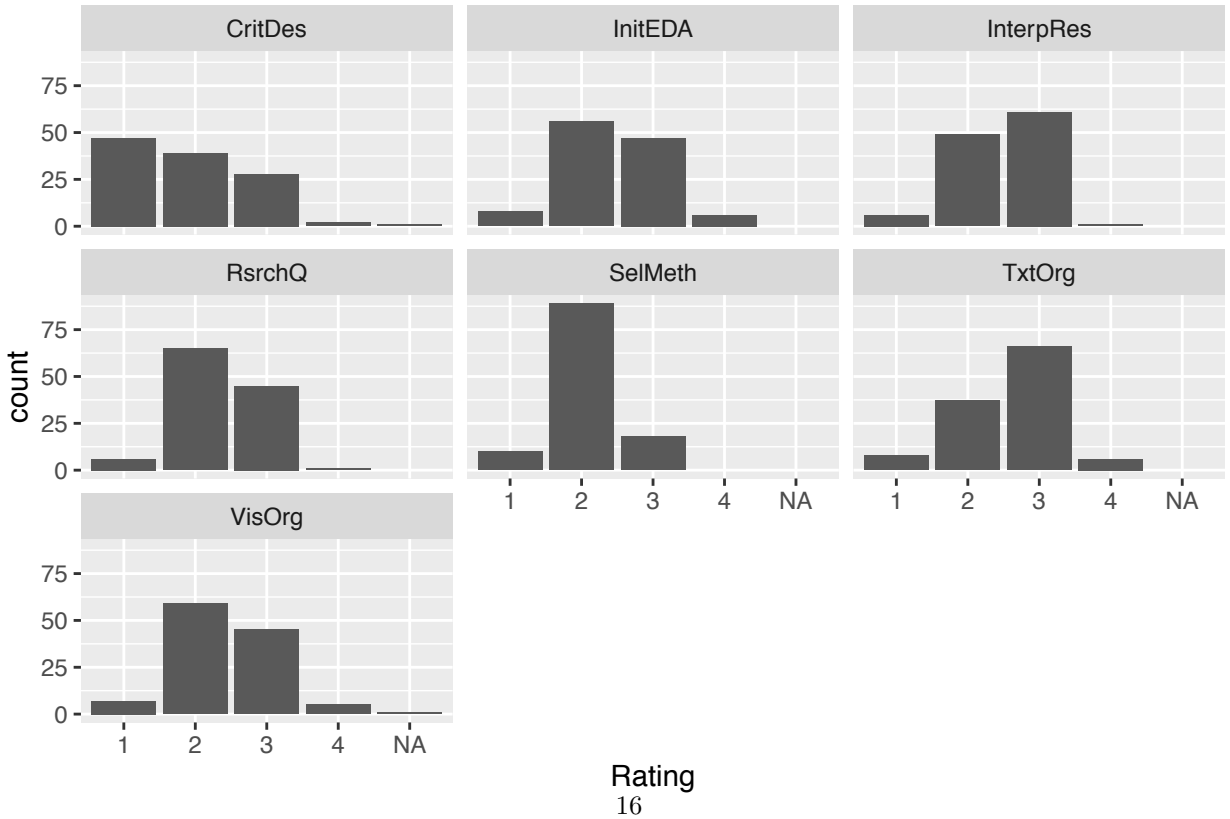
Kevin Yang

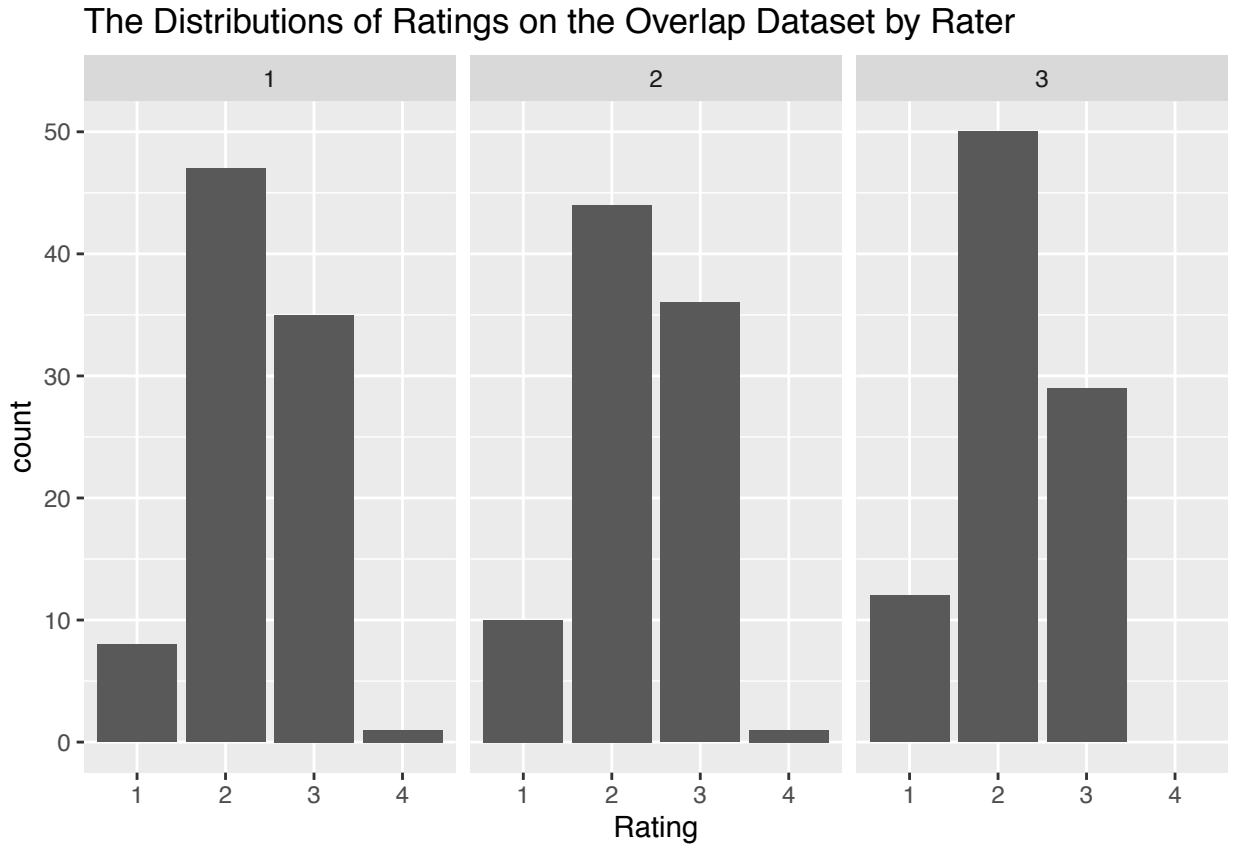
12/10/2021

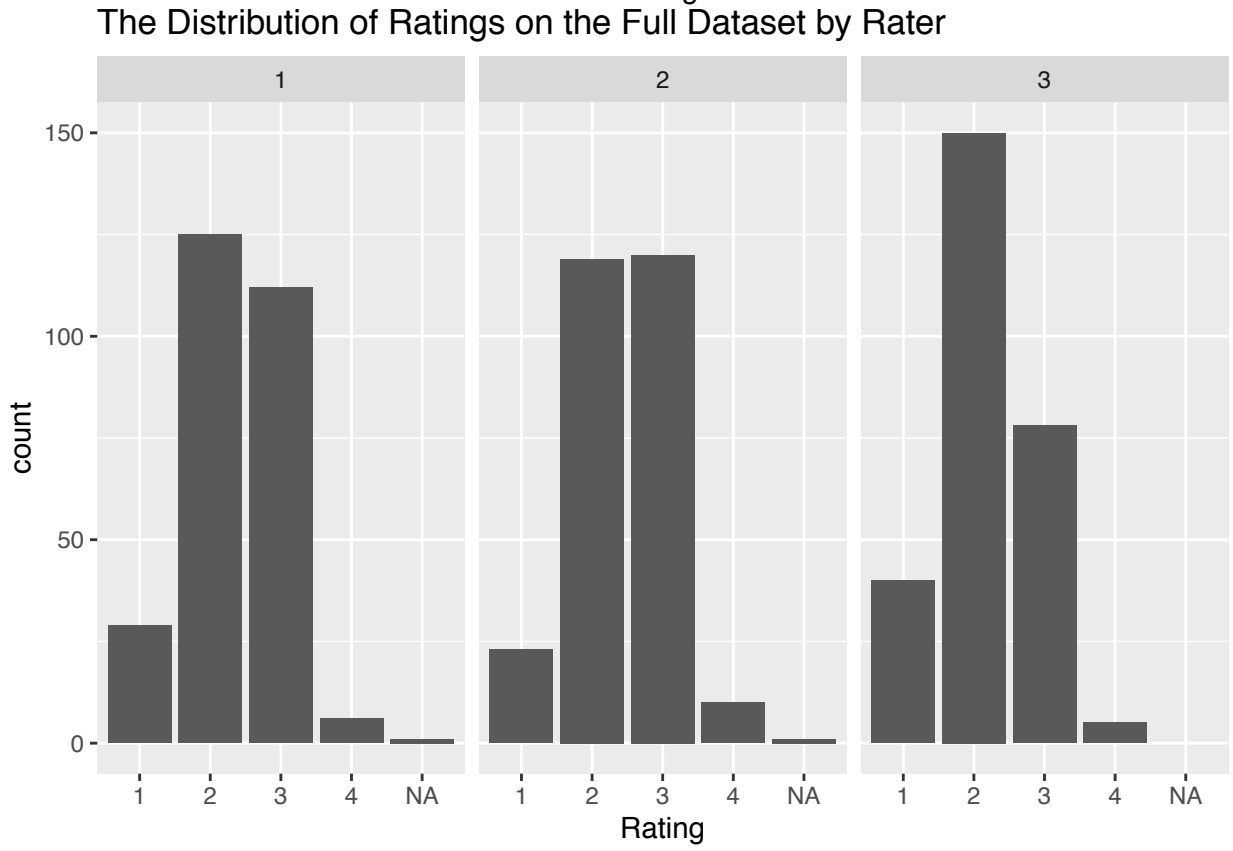
## Contents

## Appendix A: Visual Distribution of Ratings by Rubric and Rater on Both Overlapping and Full Datasets






The distribution of ratings when grouped by Rater are very similar to each other. On the full

dataset, Rater 3 does tend to give more 2's and fewer 3's than Raters 1 and 2 but they all have relatively normal distributions. On the other hand, the distribution of ratings when grouped by Rubric are wildly different from each other. CritDes for example tends to have lower scores compared to the other Rubrics with a mode of 1, InterpRes has a nearly equal amount of 2's and 3's, SelMeth has a mode of 2, and TxtOrg has a mode of 3. InitEDA, RsrchQ, and VisOrg appear to have the most similar distributions with a mode of 2, half as many 3's, and a few 1's and 4's. These distributions should be the standard for all seven Rubrics to follow.

# Appendix B: Distributions of ratings

Table 1: Mean ratings by Rater in Overlapping Artifacts

| Rater | Mean RsrchQ | Mean CritDes | Mean InitEDA | Mean SelMeth | Mean InterpRes | Mean VisOrg | Mean TxtOrg |
|---|---|---|---|---|---|---|---|
| 1 | 2.384615 | 1.615385 | 2.538461 | 2.153846 | 2.615385 | 2.153846 | 2.769231 |
| 2 | 2.153846 | 1.846154 | 2.384615 | 2.076923 | 2.615385 | 2.461539 | 2.615385 |
| 3 | 2.307692 | 1.692308 | 2.230769 | 1.923077 | 2.307692 | 2.230769 | 2.615385 |

Table 2: Mean ratings by Rater in full dataset

| Rater | Mean RsrchQ | Mean CritDes | Mean InitEDA | Mean SelMeth | Mean InterpRes | Mean VisOrg | Mean TxtOrg |
|---|---|---|---|---|---|---|---|
| 1 | 2.447368 | 1.552632 | 2.421053 | 2.105263 | 2.710526 | 2.394737 | 2.789474 |
| 2 | 2.368421 | 2.131579 | 2.578947 | 2.131579 | 2.605263 | 2.657895 | 2.578947 |
| 3 | 2.256410 | 1.897436 | 2.333333 | 1.948718 | 2.153846 | 2.205128 | 2.435897 |

Table 3: Standard Deviation of ratings by Rater in Overlapping Artifacts

| Rater | SD of RsrchQ | SD of CritDes | SD of InitEDA | SD of SelMeth | SD of InterpRes | SD of VisOrg | SD of TxtOrg |
|---|---|---|---|---|---|---|---|
| 1 | 0.5063697 | 0.6504436 | 0.6602253 | 0.3755338 | 0.5063697 | 0.5547002 | 0.5991447 |
| 2 | 0.6887372 | 0.8006408 | 0.5063697 | 0.4935481 | 0.6504436 | 0.6602253 | 0.6504436 |
| 3 | 0.4803845 | 0.7510676 | 0.4385290 | 0.6405126 | 0.6304252 | 0.5991447 | 0.6504436 |

Table 4: Standard Deviaiton of ratings by Rater in full dataset

| Rater | SD of RsrchQ | SD of CritDes | SD of InitEDA | SD of SelMeth | SD of InterpRes | SD of VisOrg | SD of TxtOrg |
|---|---|---|---|---|---|---|---|
| 1 | 0.6450380 | 0.6856588 | 0.7215441 | 0.3110117 | 0.4596059 | 0.6383879 | 0.5769395 |
| 2 | 0.6333545 | 0.9055699 | 0.6830606 | 0.4748287 | 0.5945461 | 0.6688561 | 0.7215441 |
| 3 | 0.4983102 | 0.8206182 | 0.7008766 | 0.6047495 | 0.6298898 | 0.6561245 | 0.7537580 |

Looking at these tables, it can be hard to deduce any obvious patterns, but some small observations from these tables are: Rater 3 tends to give lower ratings on average, Rater 1 has a smaller spread of ratings in most of the Rubrics, and the CritDes Rubric has a lower average and larger spread of ratings compared to the other Rubrics.

# Appendix C: How Percent Exact Agreement is Calculated

Table 5: Percent Exact Agreement Matrix, Raters 1 vs 2 on RsrchQ Rubric

| | 1 | 2 | 3 |
|---|---|---|---|
| 2 | 1 | 4 | 3 |
| 3 | 1 | 3 | 1 |

Percent Exact Agreement is calculated using matrices like the one shown above. The rows show the ratings given by Rater 1 and the columns show the ratings given by Rater 2, both for the RsrchQ Rubric. Then, the entries in the matrix where the row header equals the column header are the artifacts that received the same rating by both Raters, they are in agreement. In this case, 4 artifacts got scores of 2 and 1 artifact got a score of 3 from both Raters so they are in agreement for 5 out of the 13 artifacts in this Rubric. The Percent Exact Agreement then is 5/13 or 38.5%. This is repeated 20 more times for each pair of Raters on all seven Rubrics (Not shown to save space).

# Appendix D: ICC and agreement percentages for each Rubric and Raters

Table 6: ICC and Percent Agreement for each Rubric and Pair of Raters

| Rubric | ICC for Overlaps | ICC for Full | Rater 1 & 2 | Rater 1 & 3 | Rater 2 & 3 |
|---|---|---|---|---|---|
| RsrchQ | 0.19 | 0.21 | 0.38 | 0.77 | 0.54 |
| CritDes | 0.57 | 0.67 | 0.54 | 0.62 | 0.69 |
| InitEDA | 0.49 | 0.69 | 0.69 | 0.54 | 0.85 |
| SelMeth | 0.52 | 0.47 | 0.92 | 0.62 | 0.69 |
| InterpRes | 0.23 | 0.22 | 0.62 | 0.54 | 0.62 |
| VisOrg | 0.59 | 0.66 | 0.54 | 0.77 | 0.77 |
| TxtOrg | 0.14 | 0.19 | 0.69 | 0.62 | 0.54 |

Looking at this table, it's clear that the Raters are in agreement most of the time as their Percent Exact Agreements are usually around 50%-92% with the exception of Raters 1 & 2 on RsrchQ at 38%. For the ICCs on the other hand, the percent agreements are quite low, never going above 70% and going as low as 14% so there is a lot of discrepancies in the ratings by Rubric.

# Appendix E: Fitting All Fixed Effect Combinations to Rating on Overlapping Dataset

Table 7: ANOVA for all Rubrics on overlapping Artifacts

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| ae | 3 | 527.2803 | 538.1087 | -260.6401 | 521.2803 | NA | NA | NA |
| ab | 4 | 527.1156 | 541.5535 | -259.5578 | 519.1156 | 2.1646749 | 1 | 0.1412145 |
| ac | 4 | 528.2595 | 542.6974 | -260.1297 | 520.2595 | 0.0000000 | 0 | NA |
| ad | 4 | 528.9539 | 543.3918 | -260.4769 | 520.9539 | 0.0000000 | 0 | NA |

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| ah | 4 | 527.1156 | 541.5535 | -259.5578 | 519.1156 | 1.8382962 | 0 | NA |
| aj | 4 | 528.2595 | 542.6974 | -260.1297 | 520.2595 | 0.0000000 | 0 | NA |
| ak | 4 | 528.9539 | 543.3918 | -260.4769 | 520.9539 | 0.0000000 | 0 | NA |
| af | 5 | 528.0948 | 546.1422 | -259.0474 | 518.0948 | 2.8590776 | 1 | 0.0908596 |
| ag | 5 | 528.7892 | 546.8366 | -259.3946 | 518.7892 | 0.0000000 | 0 | NA |
| ai | 5 | 529.6953 | 547.7427 | -259.8477 | 519.6953 | 0.0000000 | 0 | NA |
| al | 5 | 529.6953 | 547.7427 | -259.8477 | 519.6953 | 0.0000000 | 0 | NA |
| am | 5 | 528.7892 | 546.8366 | -259.3946 | 518.7892 | 0.9061206 | 0 | NA |
| an | 5 | 528.0948 | 546.1422 | -259.0474 | 518.0948 | 0.6944027 | 0 | NA |
| aa | 6 | 529.5307 | 551.1875 | -258.7653 | 517.5307 | 0.5641515 | 1 | 0.4525923 |
| ao | 6 | 529.5307 | 551.1875 | -258.7653 | 517.5307 | 0.0000000 | 0 | NA |

Looking at this ANOVA table detailing the significance of every possible combination of fixed effects on a random intercept, the best model here appears to be model "ae" which contains Repeated as a fixed effect because it has the lowest BIC. But this is useless because the Overlapping dataset is being used here and Repeated is always 1 so it really shouldn't count. The next best models are models "ab" and "ah" which have Rater only and Rater and Repeated respectively as fixed effects. Since we ruled Repeated as an invalid effect, this leaves the Rater only model as the best model here.

## Appendix F: Summary of Rater only Model from previous Appendix

```
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: Rating ~ 1 + Rater + (1 | Artifact)
#>    Data: talloverlap
#>
#> REML criterion at convergence: 526.7
#>
#> Scaled residuals:
#>     Min      1Q  Median      3Q     Max
#> -2.6754 -0.6404 -0.0417  0.8514  3.1122
#>
#> Random effects:
#>  Groups   Name        Variance Std.Dev.
#>  Artifact (Intercept) 0.07194  0.2682
#>  Residual             0.36540  0.6045
#> Number of obs: 273, groups:  Artifact, 13
#>
#> Fixed effects:
#>             Estimate Std. Error t value
#> (Intercept)  2.40293    0.12208  19.684
#> Rater       -0.06593    0.04481  -1.472
#>
#> Correlation of Fixed Effects:
#>       (Intr)
#> Rater -0.734
```

However, looking at the summary statistics for the Rater only model from Appendix E, it looks like Rater is not significant either because it has a t-value of -1.472. The fixed effects can only be significant if the t-value is less than -2 or greater than 2, which it is not here. As such it is likely that the random intercept model is the best model to predict an artifact's rating when disregarding Rubrics.

# Appendix G: Fitting All Fixed Effect Combinations to Rating on Full Dataset

Table 8: ANOVA for all Rubrics on all artifacts

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| bb | 4 | 1641.161 | 1659.984 | -816.5807 | 1633.161 | NA | NA | NA |
| bc | 4 | 1644.243 | 1663.065 | -818.1213 | 1636.243 | 0.0000000 | 0 | NA |
| be | 4 | 1645.454 | 1664.277 | -818.7271 | 1637.454 | 0.0000000 | 0 | NA |
| bd | 5 | 1645.278 | 1668.806 | -817.6389 | 1635.278 | 2.1762592 | 1 | 0.1401548 |
| bf | 5 | 1641.429 | 1664.957 | -815.7145 | 1631.429 | 3.8487807 | 0 | NA |
| bh | 5 | 1642.785 | 1666.313 | -816.3924 | 1632.785 | 0.0000000 | 0 | NA |
| bj | 5 | 1645.746 | 1669.274 | -817.8729 | 1635.746 | 0.0000000 | 0 | NA |
| bg | 6 | 1641.830 | 1670.063 | -814.9148 | 1629.830 | 5.9162093 | 1 | 0.0150022 |
| bi | 6 | 1645.940 | 1674.174 | -816.9702 | 1633.940 | 0.0000000 | 0 | NA |
| bk | 6 | 1646.984 | 1675.218 | -817.4922 | 1634.984 | 0.0000000 | 0 | NA |
| bn | 6 | 1642.909 | 1671.143 | -815.4547 | 1630.909 | 4.0750460 | 0 | NA |
| bl | 7 | 1647.534 | 1680.474 | -816.7671 | 1633.534 | 0.0000000 | 1 | 1.0000000 |
| bm | 7 | 1643.532 | 1676.471 | -814.7658 | 1629.532 | 4.0025326 | 0 | NA |
| bo | 7 | 1642.435 | 1675.375 | -814.2175 | 1628.435 | 1.0964901 | 0 | NA |
| ba | 8 | 1644.020 | 1681.665 | -814.0101 | 1628.020 | 0.4149393 | 1 | 0.5194731 |

Looking at the ANOVA table it looks like there could be two good models here: model "bb" which is Rater only because it has the lowest AIC and BIC, and model "bg" which is Rater and Sex because it is the only significant model following the Chi-Squared column.

# Appendix H: Summary of Rater only Model from Appendix G

```
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: Rating ~ 1 + Rater + (1 | Artifact)
#>    Data: tall
#>
#> REML criterion at convergence: 1642.4
#>
#> Scaled residuals:
#>     Min      1Q  Median      3Q     Max
#> -2.7220 -0.5998 -0.0295  0.7807  3.0839
#>
#> Random effects:
#>  Groups   Name        Variance Std.Dev.
#>  Artifact (Intercept) 0.1288   0.3589
#>  Residual             0.3726   0.6104
#> Number of obs: 817, groups:  Artifact, 91
#>
#> Fixed effects:
#>             Estimate Std. Error t value
#> (Intercept)  2.48489    0.08431  29.474
#> Rater       -0.07756    0.03595  -2.158
#>
#> Correlation of Fixed Effects:
#>       (Intr)
#> Rater -0.853
```

According to these summary statistics, Rater is significant in predicting rating on the full dataset since it's t-value is less than -2.

# Appendix I: Summary of Rater and Sex Model from Appendix G

```
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: Rating ~ 1 + Rater + Sex + (1 | Artifact)
#>    Data: tall
#>
#> REML criterion at convergence: 1642
#>
#> Scaled residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.73227 -0.60863 -0.03954  0.77540  3.07143
#>
#> Random effects:
#>  Groups   Name        Variance Std.Dev.
#>  Artifact (Intercept) 0.1263   0.3554
#>  Residual             0.3726   0.6104
#> Number of obs: 817, groups:  Artifact, 91
#>
#> Fixed effects:
#>             Estimate Std. Error t value
#> (Intercept)  3.25078    0.43732   7.433
#> Rater       -0.08359    0.03602  -2.321
#> SexF        -0.77417    0.42953  -1.802
#> SexM        -0.74674    0.43020  -1.736
#>
#> Correlation of Fixed Effects:
#>       (Intr) Rater  SexF
#> Rater -0.247
#> SexF  -0.978  0.089
#> SexM  -0.974  0.080  0.979
```

According to these summary statistics, Rater is once again significant with a t-value less than -2 but Sex is not significant since both factors have t-values between -2 and 2. This means that model "bb" from Appendix G with Rater only as a fixed effect is the best model for predicting ratings on the full dataset disregarding Rubrics.

# Appendix J: Formulas for the Fixed Effects Models in Overlapping Artifacts

```
#> $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)
```

Modeling each Rubric separately, this output suggests models identical to the conclusion Appendix F which is that a random intercept model is the best model to predict ratings on the overlapping dataset.

## Appendix K: Formulas for the Fixed Effects Models in All Artifacts

```
#> $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
```

This output shows that on the full dataset, certain effects have a higher significance in predicting ratings on a select number of Rubrics. For CritDes, InterpRes, and VisOrg, the Rater has a significant effect on the rating, and for SelMeth, Rater and Semester have a significant effect on the rating

## Appendix L: Coefficients and Significance of Random Effects/Interaction Terms/Random Intercepts

For each of the Rubrics, summary statistics are provided below showing the significance of the random intercepts and fixed effects. For SelMeth specifically, an additional significance test was conducted to see if an interaction term was needed between Rater and Semester. Significance is determined if a t-value is less than -2 or greater than 2, or if a p-value is less than 0.05.

### SelMeth

Table 9: Significance of random effects terms for SelMeth Rubric

| | 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 |

All random effects terms are significant here.

Table 10: Significance of the Rater intercept term for SelMeth

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| tmp.single_intercept | 4 | 145.0688 | 156.0832 | -68.53441 | 137.0688 | NA | NA | NA |
| tmp | 6 | 142.0543 | 158.5758 | -65.02713 | 130.0543 | 7.014565 | 2 | 0.0299783 |

Intercept is significant.

Table 11: ANOVA for the interaction terms

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| tmp | 6 | 142.0543 | 158.5758 | -65.02713 | 130.0543 | NA | NA | NA |
| tmp.fixed_interactions | 8 | 143.4622 | 165.4910 | -63.73112 | 127.4622 | 2.592023 | 2 | 0.2736209 |

Chi-Squared p-value is greater than 0.05 so the interaction term is not significant and not necessary for this model.

## SelMeth Summary

```
#> 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
```

## CritDes

Table 12: Significance of random effects terms for CritDes Rubric

| | 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 |

The Rater random effect term is significant here.

Table 13: Significance of the Rater intercept term for CritDes

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| tmp.single_intercept | 3 | 277.6769 | 285.9116 | -135.8384 | 271.6769 | NA | NA | NA |
| tmp | 5 | 273.6233 | 287.3480 | -131.8117 | 263.6233 | 8.05352 | 2 | 0.017832 |

Intercept is significant too.

## CritDes Summary

```
#> 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
```

## InterpRes

Table 14: Significance of random effects terms for InterpRes Rubric

| | Estimate | Std. Error | t value |
|---|---|---|---|
| as.factor(Rater)1 | 2.70 | 0.09 | 30.34 |
| as.factor(Rater)2 | 2.59 | 0.09 | 29.01 |
| as.factor(Rater)3 | 2.14 | 0.09 | 23.70 |

The Rater random effect term is significant here.

Table 15: Significance of the Rater intercept term for InterpRes

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| tmp.single_intercept | 3 | 218.5257 | 226.7865 | -106.26287 | 212.5257 | NA | NA | NA |
| tmp | 5 | 200.6614 | 214.4294 | -95.33072 | 190.6614 | 21.86429 | 2 | 1.79e-05 |

Intercept is significant too.

## InterpRes Summary

```
#> 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
```

## VisOrg

Table 16: Significance of random effects terms for VisOrg Rubric

| | 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 |

The Rater random effect term is significant here.

Table 17: Significance of the Rater intercept term for VisOrg

| | npar | AIC | BIC | logLik | deviance | Chisq | Df | Pr(>Chisq) |
|---|---|---|---|---|---|---|---|---|
| tmp.single_intercept | 3 | 227.2078 | 235.4426 | -110.6039 | 221.2078 | NA | NA | NA |
| tmp | 5 | 220.8158 | 234.5404 | -105.4079 | 210.8158 | 10.39204 | 2 | 0.0055386 |

Intercept is significant too.

## VisOrg Summary

```
#> 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
```

## InitEDA

```
fla <- formula(model.formula.alldata[["InitEDA"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "InitEDA", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for InitEDA Rubric")
```

Table 18: Estimate and Significance of random intercept term for InitEDA Rubric

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 2.44 | 0.08 | 32.4 |

Random intercept is significant.

## InitEDA Summary

```
summary(tmp)
```

```
#> 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
```

## RsrchQ

```
fla <- formula(model.formula.alldata[["RsrchQ"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "RsrchQ", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for RsrchQ Rubric")
```

Table 19: Estimate and Significance of random intercept term for RsrchQ Rubric

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 2.35 | 0.06 | 40.59 |

Random intercept is significant.

## RsrchQ summary

```
summary(tmp)
```

```
#> 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
```

## TxtOrg

```
fla <- formula(model.formula.alldata[["TxtOrg"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "TxtOrg", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for TxtOrg Rubric")
```

Table 20: Estimate and Significance of random intercept term for TxtOrg Rubric

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 2.59 | 0.07 | 37.93 |

Random intercept is significant.

## TxtOrg summary

```
summary(tmp)
```

```
#> 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
```

# Appendix M: Table of all coefficients and standard errors

Table 21: Coefficients for the Final Models for each Rubric

| | RsrchQ | CritDes | InitEDA | SelMeth | InterpRes | VisOrg | TxtOrg |
|---|---|---|---|---|---|---|---|
| Intercept | 2.35 | 0.00 | 2.44 | 0.00 | 0.00 | 0.00 | 2.59 |
| Rater 1 | 0.00 | 1.69 | 0.00 | 2.25 | 2.70 | 2.38 | 0.00 |
| Rater 2 | 0.00 | 2.11 | 0.00 | 2.23 | 2.59 | 2.65 | 0.00 |
| Rater 3 | 0.00 | 1.89 | 0.00 | 2.03 | 2.14 | 2.28 | 0.00 |
| SemesterS19 | 0.00 | 0.00 | 0.00 | -0.36 | 0.00 | 0.00 | 0.00 |
| Variance by Artifact | 0.07 | 0.43 | 0.37 | 0.09 | 0.06 | 0.29 | 0.09 |
| Variance of Residuals | 0.28 | 0.24 | 0.17 | 0.11 | 0.25 | 0.15 | 0.40 |

Here is a table showcasing all of the coefficients needed for predicting ratings for each Rubric. The variances on the bottom two rows of the table "randomness" of the models, how far off the intercept or fixed effects a prediction can be. Variance by Artifact covers the "randomness" of the model after grouping the data by Artifact. Variance by Residual covers the "randomness" of the model as a whole with no grouping.

# Appendix N: Significant Fixed and Interaction Terms after making a model with all possible Combinations

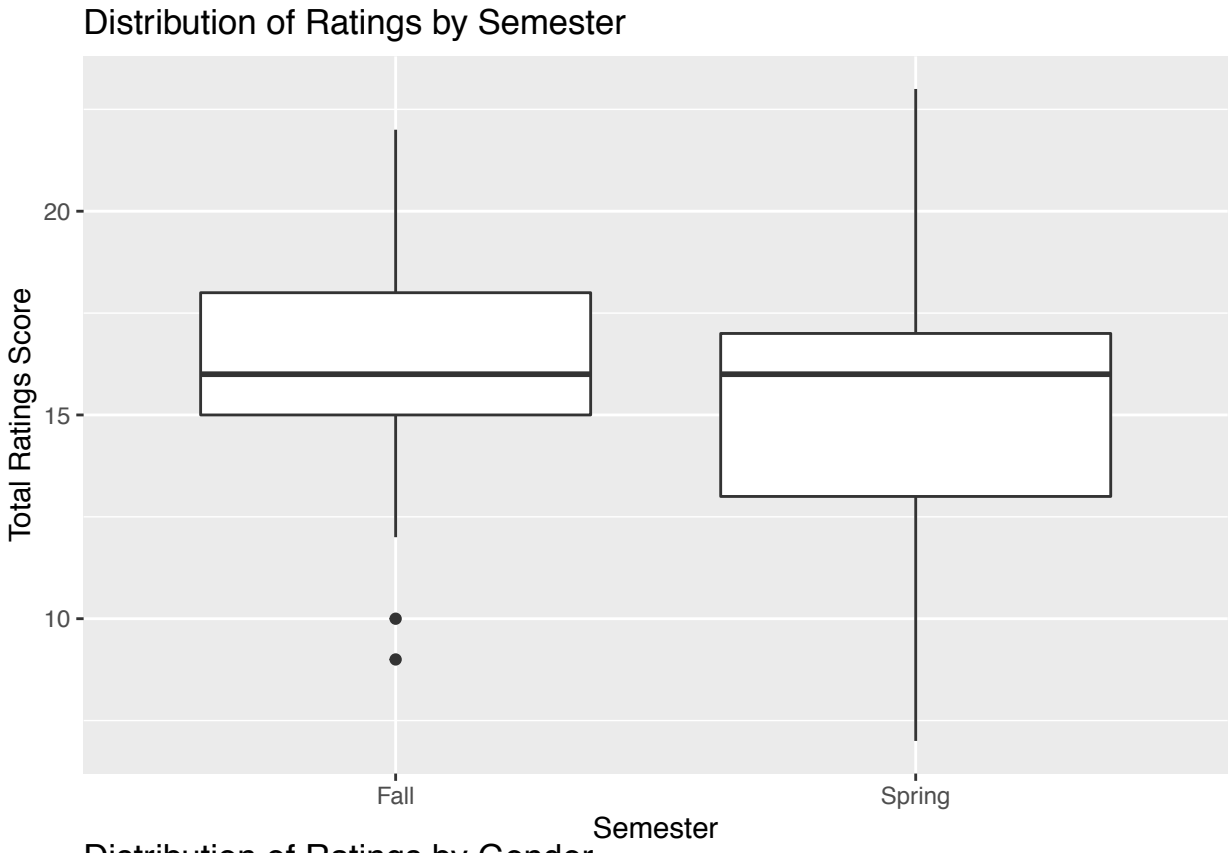
Table 22: Significant Fixed Effects and Interaction Terms

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 1.62 | 0.29 | 5.63 |
| RubricInitEDA | 0.88 | 0.33 | 2.69 |
| RubricInterpRes | 1.25 | 0.31 | 4.05 |
| RubricRsrchQ | 0.75 | 0.28 | 2.67 |
| RubricTxtOrg | 1.25 | 0.28 | 4.42 |
| RubricVisOrg | 1.19 | 0.32 | 3.74 |
| as.factor(Rater)2:RubricInitEDA | -1.00 | 0.46 | -2.17 |
| as.factor(Rater)2:RubricInterpRes | -1.13 | 0.44 | -2.58 |
| as.factor(Rater)3:RubricInterpRes | -1.11 | 0.45 | -2.45 |
| as.factor(Rater)2:RubricTxtOrg | -1.25 | 0.40 | -3.13 |
| as.factor(Rater)3:RubricVisOrg | -1.04 | 0.46 | -2.28 |
| as.factor(Rater)2:SexM:RubricInitEDA | 1.44 | 0.63 | 2.31 |

| | Estimate | Std. Error | t value |
|---|---|---|---|
| as.factor(Rater)2:SexM:RubricTxtOrg | 1.46 | 0.54 | 2.69 |
| as.factor(Rater)2:SexM:RubricVisOrg | 1.25 | 0.60 | 2.10 |
| as.factor(Rater)2:Repeated:RubricVisOrg | 1.21 | 0.60 | 2.01 |
| as.factor(Rater)2:SexM:Repeated:RubricTxtOrg | -1.66 | 0.79 | -2.09 |
| as.factor(Rater)2:SexM:Repeated:RubricVisOrg | -1.85 | 0.83 | -2.23 |

Purely for exploratory purposes one giant linear model was created containing every fixed effect and every combination of interaction terms. As there are many interaction terms, only the significant interaction terms are in the table above. These show the change in an artifact's rating in one factor combination compared to an artifact's rating in a similar factor combination (Male vs female, Fall vs Spring)

## Appendix O: Distribution of Ratings Based on Gender and Semester



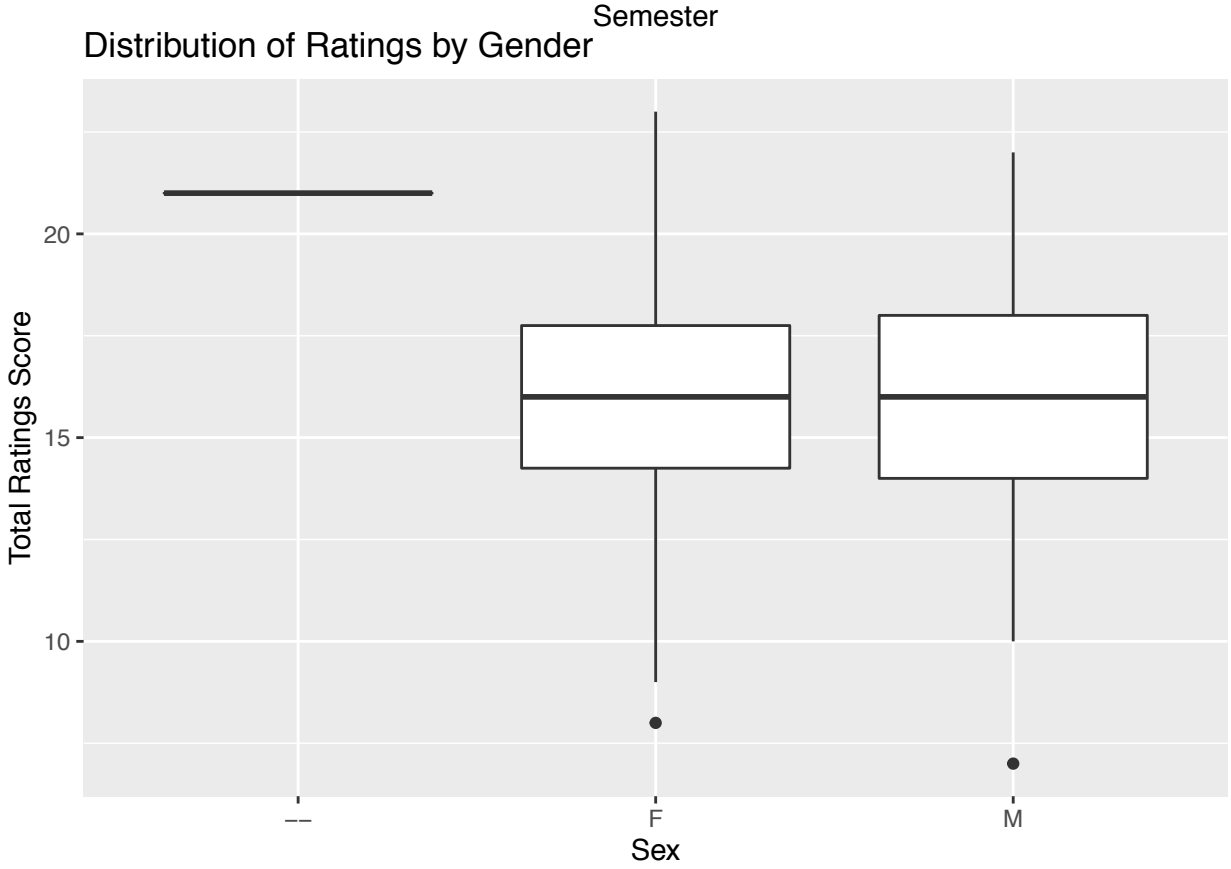



Table 23: Average Rating of Artifacts by Gender

| Sex | genderrating | gendermean | count |
|---|---|---|---|
| – | 21 | 21.00000 | 1 |
| F | 1004 | 16.19355 | 62 |
| M | 841 | 16.17308 | 52 |

Table 24: Average Rating of Artifacts by Semester

| Semester | semesterrating | semestermean | count |
|---|---|---|---|
| Fall | 1351 | 16.47561 | 82 |
| Spring | 515 | 15.60606 | 33 |

Lastly, some boxplots were made to visualize the distribution of ratings when grouped by gender or semester. Using the sum of all of the Rubric ratings as the response, the boxplots show that the distribution of ratings when grouped by gender is very similar in size and shape while grouping by semester results in a smaller spread in the fall and a larger spread in the spring.

The tables give some more detail on the boxplots as the mean total rating by gender are very close to each other too, off by only 0.02 while the mean total rating by semester is more different, off by 0.87 but that could likely be because the class size is much smaller in the spring than in the fall meaning it would be harder to make the ratings more normally distributed.

# Code Appendix

```
knitr::opts_chunk$set(comment = "#>", tidy.opts = list(width.cutoff = 40),
    tidy = TRUE)
library(arm)
library(lme4)
library(plyr)
library(tidyverse)
library(performance)
library(LMERConvenienceFunctions)
library(RLRsim)
setwd("~/Documents/College/Semester 9/Applied Linear Modeling/ALM HW10")
ratings <- read.csv("ratings.csv")
tall <- read.csv("tall.csv")  # Rows 5,122,239,356,473,590,707 have NAs
tall$Sex <- as.character(tall$Sex)
tall[c(5, 122, 239, 356, 473, 590, 707),
    6] <- "--"
tall$Rating <- factor(tall$Rating, levels = 1:4)
for (i in unique(tall$Rubric)) {
    ratings[, i] <- factor(ratings[, i],
        levels = 1:4)
}
ratingsoverlap <- ratings[is.na(ratings$Overlap) ==
    FALSE, ]
talloverlap <- tall[tall$Repeated == 1, ]
par(mfrow = c(2, 2))
ggplot(talloverlap, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar() + ggtitle("The Distribution of Ratings in Overlapping Artifacts by Rubric")
```

```
ggplot(tall, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar() + ggtitle("The Distribution of Ratings in All Artifacts by Rubric")

ggplot(talloverlap, aes(x = Rating)) + facet_wrap(~Rater) +
    geom_bar() + ggtitle("The Distributions of Ratings on the Overlap Dataset by Rater")

ggplot(tall, aes(x = Rating)) + facet_wrap(~Rater) +
    geom_bar() + ggtitle("The Distribution of Ratings on the Full Dataset by Rater")
setwd("~/Documents/College/Semester 9/Applied Linear Modeling/ALM HW10")
ratings <- read.csv("ratings.csv")
ratingsoverlap <- ratings[is.na(ratings$Overlap) ==
    FALSE, ]
knitr::kable(ratingsoverlap %>%
    dplyr::group_by(Rater) %>%
    dplyr::summarize(`Mean RsrchQ` = mean(RsrchQ),
        ` Mean CritDes` = mean(CritDes),
        `Mean InitEDA` = mean(InitEDA), `Mean SelMeth` = mean(SelMeth),
        `Mean InterpRes` = mean(InterpRes),
        `Mean VisOrg` = mean(VisOrg), `Mean TxtOrg` = mean(TxtOrg)),
    caption = "Mean ratings by Rater in Overlapping Artifacts")

knitr::kable(ratings %>%
    drop_na(RsrchQ, CritDes, InitEDA, SelMeth,
        InterpRes, VisOrg, TxtOrg) %>%
    dplyr::group_by(Rater) %>%
    dplyr::summarize(`Mean RsrchQ` = mean(RsrchQ),
        ` Mean CritDes` = mean(CritDes),
        `Mean InitEDA` = mean(InitEDA), `Mean SelMeth` = mean(SelMeth),
        `Mean InterpRes` = mean(InterpRes),
        `Mean VisOrg` = mean(VisOrg), `Mean TxtOrg` = mean(TxtOrg)),
    caption = "Mean ratings by Rater in full dataset")

knitr::kable(ratingsoverlap %>%
    dplyr::group_by(Rater) %>%
    dplyr::summarize(`SD of RsrchQ` = sd(RsrchQ),
        `SD of CritDes` = sd(CritDes), `SD of InitEDA` = sd(InitEDA),
        `SD of SelMeth` = sd(SelMeth), `SD of InterpRes` = sd(InterpRes),
        `SD of VisOrg` = sd(VisOrg), `SD of TxtOrg` = sd(TxtOrg)),
    caption = "Standard Deviation of ratings by Rater in Overlapping Artifacts")

knitr::kable(ratings %>%
    drop_na(RsrchQ, CritDes, InitEDA, SelMeth,
        InterpRes, VisOrg, TxtOrg) %>%
    dplyr::group_by(Rater) %>%
    dplyr::summarize(`SD of RsrchQ` = sd(RsrchQ),
        `SD of CritDes` = sd(CritDes), `SD of InitEDA` = sd(InitEDA),
        `SD of SelMeth` = sd(SelMeth), `SD of InterpRes` = sd(InterpRes),
        `SD of VisOrg` = sd(VisOrg), `SD of TxtOrg` = sd(TxtOrg)),
    caption = "Standard Deviaiton of ratings by Rater in full dataset")
talloverlap$Rating <- as.numeric(talloverlap$Rating)
tall$Rating <- as.numeric(tall$Rating)
RsrchQ <- talloverlap[talloverlap$Rubric ==
    "RsrchQ", ]
```

```
CritDes <- talloverlap[talloverlap$Rubric ==
    "CritDes", ]
InitEDA <- talloverlap[talloverlap$Rubric ==
    "InitEDA", ]
SelMeth <- talloverlap[talloverlap$Rubric ==
    "SelMeth", ]
InterpRes <- talloverlap[talloverlap$Rubric ==
    "InterpRes", ]
VisOrg <- talloverlap[talloverlap$Rubric ==
    "VisOrg", ]
TxtOrg <- talloverlap[talloverlap$Rubric ==
    "TxtOrg", ]
a <- lmer(Rating ~ 1 + (1 | Artifact), data = RsrchQ)
b <- lmer(Rating ~ 1 + (1 | Artifact), data = CritDes)
c <- lmer(Rating ~ 1 + (1 | Artifact), data = InitEDA)
d <- lmer(Rating ~ 1 + (1 | Artifact), data = SelMeth)
e <- lmer(Rating ~ 1 + (1 | Artifact), data = InterpRes)
f <- lmer(Rating ~ 1 + (1 | Artifact), data = VisOrg)
g <- lmer(Rating ~ 1 + (1 | Artifact), data = TxtOrg)
RsrchQall <- tall[tall$Rubric == "RsrchQ",
    ]
CritDesall <- tall[tall$Rubric == "CritDes",
    ]
InitEDAall <- tall[tall$Rubric == "InitEDA",
    ]
SelMethall <- tall[tall$Rubric == "SelMeth",
    ]
InterpResall <- tall[tall$Rubric == "InterpRes",
    ]
VisOrgall <- tall[tall$Rubric == "VisOrg",
    ]
TxtOrgall <- tall[tall$Rubric == "TxtOrg",
    ]
h <- lmer(Rating ~ 1 + (1 | Artifact), data = RsrchQall)
j <- lmer(Rating ~ 1 + (1 | Artifact), data = CritDesall)
k <- lmer(Rating ~ 1 + (1 | Artifact), data = InitEDAall)
l <- lmer(Rating ~ 1 + (1 | Artifact), data = SelMethall)
m <- lmer(Rating ~ 1 + (1 | Artifact), data = InterpResall)
n <- lmer(Rating ~ 1 + (1 | Artifact), data = VisOrgall)
o <- lmer(Rating ~ 1 + (1 | Artifact), data = TxtOrgall)
rubricnames <- c("RsrchQ", "CritDes", "InitEDA",
    "SelMeth", "InterpRes", "VisOrg", "TxtOrg")
icc1 <- rbind(icc(a), icc(b), icc(c), icc(d),
    icc(e), icc(f), icc(g))[, 1]
icc2 <- rbind(icc(h), icc(j), icc(k), icc(l),
    icc(m), icc(n), icc(o))[, 1]
icctable <- cbind(rubricnames, icc1, icc2)
colnames(icctable) <- c("Rubric", "ICC for Overlaps",
    "Icc for Full")
knitr::kable(table(RsrchQ[RsrchQ$Rater ==
    1, 8], RsrchQ[RsrchQ$Rater == 2, 8]),
    caption = "Percent Exact Agreement Matrix, Raters 1 vs 2 on RsrchQ Rubric")  # 5 matches
table(RsrchQ[RsrchQ$Rater == 1, 8], RsrchQ[RsrchQ$Rater ==
```

```
    3, 8])  # 10 matches
table(RsrchQ[RsrchQ$Rater == 2, 8], RsrchQ[RsrchQ$Rater ==
    3, 8])  # 7 matches
table(CritDes[CritDes$Rater == 1, 8], CritDes[CritDes$Rater ==
    2, 8])  # 7 matches
table(CritDes[CritDes$Rater == 1, 8], CritDes[CritDes$Rater ==
    3, 8])  # 8 matches
table(CritDes[CritDes$Rater == 2, 8], CritDes[CritDes$Rater ==
    3, 8])  # 9 matches
table(InitEDA[InitEDA$Rater == 1, 8], InitEDA[InitEDA$Rater ==
    2, 8])  # 9 matches
table(InitEDA[InitEDA$Rater == 1, 8], InitEDA[InitEDA$Rater ==
    3, 8])  # 7 matches
table(InitEDA[InitEDA$Rater == 2, 8], InitEDA[InitEDA$Rater ==
    3, 8])  # 11 matches
table(SelMeth[SelMeth$Rater == 1, 8], SelMeth[SelMeth$Rater ==
    2, 8])  # 12 matches
table(SelMeth[SelMeth$Rater == 1, 8], SelMeth[SelMeth$Rater ==
    3, 8])  # 8 matches
table(SelMeth[SelMeth$Rater == 2, 8], SelMeth[SelMeth$Rater ==
    3, 8])  # 9 matches
table(InterpRes[InterpRes$Rater == 1, 8],
    InterpRes[InterpRes$Rater == 2, 8])  # 8 matches
table(InterpRes[InterpRes$Rater == 1, 8],
    InterpRes[InterpRes$Rater == 3, 8])  # 7 matches
table(InterpRes[InterpRes$Rater == 2, 8],
    InterpRes[InterpRes$Rater == 3, 8])  # 8 matches
table(VisOrg[VisOrg$Rater == 1, 8], VisOrg[VisOrg$Rater ==
    2, 8])  # 7 matches
table(VisOrg[VisOrg$Rater == 1, 8], VisOrg[VisOrg$Rater ==
    3, 8])  # 10 matches
table(VisOrg[VisOrg$Rater == 2, 8], VisOrg[VisOrg$Rater ==
    3, 8])  # 10 matches
table(TxtOrg[TxtOrg$Rater == 1, 8], TxtOrg[TxtOrg$Rater ==
    2, 8])  # 9 matches
table(TxtOrg[TxtOrg$Rater == 1, 8], TxtOrg[TxtOrg$Rater ==
    3, 8])  # 8 matches
table(TxtOrg[TxtOrg$Rater == 2, 8], TxtOrg[TxtOrg$Rater ==
    3, 8])  # 7 matches
r12 <- c(5/13, 7/13, 9/13, 12/13, 8/13, 7/13,
    9/13)
r13 <- c(10/13, 8/13, 7/13, 8/13, 7/13, 10/13,
    8/13)
r23 <- c(7/13, 9/13, 11/13, 9/13, 8/13, 10/13,
    7/13)
icctable <- as.data.frame(cbind(rubricnames,
    as.numeric(icc1), as.numeric(icc2), as.numeric(r12),
    as.numeric(r13), as.numeric(r23)))
icctable$V2 <- as.numeric(as.character(icctable$V2))
icctable$V3 <- as.numeric(as.character(icctable$V3))
icctable$V4 <- as.numeric(as.character(icctable$V4))
icctable$V5 <- as.numeric(as.character(icctable$V5))
icctable$V6 <- as.numeric(as.character(icctable$V6))
```

```
colnames(icctable) <- c("Rubric", "ICC for Overlaps",
    "ICC for Full", "Rater 1 & 2", "Rater 1 & 3",
    "Rater 2 & 3")
options(digits = 2)
knitr::kable(icctable, caption = "ICC and Percent Agreement for each Rubric and Pair of Raters")
options(digits = 7)
aa <- lmer(Rating ~ 1 + Rater + Semester +
    Sex + Repeated + (1 | Artifact), data = talloverlap)
ab <- lmer(Rating ~ 1 + Rater + (1 | Artifact),
    data = talloverlap)
ac <- lmer(Rating ~ 1 + Semester + (1 | Artifact),
    data = talloverlap)
ad <- lmer(Rating ~ 1 + Sex + (1 | Artifact),
    data = talloverlap)
ae <- lmer(Rating ~ 1 + Repeated + (1 | Artifact),
    data = talloverlap)
af <- lmer(Rating ~ 1 + Rater + Semester +
    (1 | Artifact), data = talloverlap)
ag <- lmer(Rating ~ 1 + Rater + Sex + (1 |
    Artifact), data = talloverlap)
ah <- lmer(Rating ~ 1 + Rater + Repeated +
    (1 | Artifact), data = talloverlap)
ai <- lmer(Rating ~ 1 + Semester + Sex +
    (1 | Artifact), data = talloverlap)
aj <- lmer(Rating ~ 1 + Semester + Repeated +
    (1 | Artifact), data = talloverlap)
ak <- lmer(Rating ~ 1 + Sex + Repeated +
    (1 | Artifact), data = talloverlap)
al <- lmer(Rating ~ 1 + Semester + Sex +
    Repeated + (1 | Artifact), data = talloverlap)
am <- lmer(Rating ~ 1 + Rater + Sex + Repeated +
    (1 | Artifact), data = talloverlap)
an <- lmer(Rating ~ 1 + Rater + Semester +
    Repeated + (1 | Artifact), data = talloverlap)
ao <- lmer(Rating ~ 1 + Rater + Semester +
    Sex + (1 | Artifact), data = talloverlap)
knitr::kable(anova(aa, ab, ac, ad, ae, af,
    ag, ah, ai, aj, ak, al, am, an, ao),
    caption = "ANOVA for all Rubrics on overlapping Artifacts")
# ae (Repeated only) has lowest BIC,
# but since this model is overlap only,
# we won't count it Next best models
# are ab and ah (Rater only and Rater
# and Repeated), removing repeated
# leaves Rater only
summary(ab)
ba <- lmer(Rating ~ 1 + Rater + Semester +
    Sex + Repeated + (1 | Artifact), data = tall)
bb <- lmer(Rating ~ 1 + Rater + (1 | Artifact),
    data = tall)
bc <- lmer(Rating ~ 1 + Semester + (1 | Artifact),
    data = tall)
bd <- lmer(Rating ~ 1 + Sex + (1 | Artifact),
```

```
    data = tall)
be <- lmer(Rating ~ 1 + Repeated + (1 | Artifact),
    data = tall)
bf <- lmer(Rating ~ 1 + Rater + Semester +
    (1 | Artifact), data = tall)
bg <- lmer(Rating ~ 1 + Rater + Sex + (1 |
    Artifact), data = tall)
bh <- lmer(Rating ~ 1 + Rater + Repeated +
    (1 | Artifact), data = tall)
bi <- lmer(Rating ~ 1 + Semester + Sex +
    (1 | Artifact), data = tall)
bj <- lmer(Rating ~ 1 + Semester + Repeated +
    (1 | Artifact), data = tall)
bk <- lmer(Rating ~ 1 + Sex + Repeated +
    (1 | Artifact), data = tall)
bl <- lmer(Rating ~ 1 + Semester + Sex +
    Repeated + (1 | Artifact), data = tall)
bm <- lmer(Rating ~ 1 + Rater + Sex + Repeated +
    (1 | Artifact), data = tall)
bn <- lmer(Rating ~ 1 + Rater + Semester +
    Repeated + (1 | Artifact), data = tall)
bo <- lmer(Rating ~ 1 + Rater + Semester +
    Sex + (1 | Artifact), data = tall)
knitr::kable(anova(ba, bb, bc, bd, be, bf,
    bg, bh, bi, bj, bk, bl, bm, bn, bo),
    caption = "ANOVA for all Rubrics on all artifacts")
## bb (Rater only) has lowest AIC and
## BIC, while bg (Rater and Sex) is the
## only significant model
summary(bb)
summary(bg)
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) {

    rubric.data <- talloverlap[talloverlap$Rubric ==
        i, ]
    tmp <- lmer(as.numeric(Rating) ~ -1 +
        as.factor(Rater) + Semester + Sex +
        (1 | Artifact), data = rubric.data,
        REML = FALSE)

    tmp.back_elim <- fitLMER.fnc(tmp, set.REML.FALSE = TRUE,
        log.file.name = FALSE)

    tmp.single_intercept <- update(tmp.back_elim,
        . ~ . + 1 - as.factor(Rater))
    pval <- anova(tmp.single_intercept, tmp.back_elim)$"Pr(>Chisq)"[2]

    if (pval <= 0.05) {
        tmp_final <- tmp.back_elim
    } else {
```

```
        tmp_final <- tmp.single_intercept
    }

    model.formula.13[[i]] <- formula(tmp_final)

}
model.formula.13
Rubric.names <- sort(unique(tall$Rubric))
tall.nonmissing <- tall[-c(161, 684), ]
tall.nonmissing <- tall.nonmissing[tall.nonmissing$Sex !=
    "--", ]
model.formula.alldata <- as.list(rep(NA,
    7))
names(model.formula.alldata) <- Rubric.names
for (i in Rubric.names) {

    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)

    tmp.back_elim <- fitLMER.fnc(tmp, set.REML.FALSE = TRUE,
        log.file.name = FALSE)

    tmp.single_intercept <- update(tmp.back_elim,
        . ~ . + 1 - as.factor(Rater))
    pval <- anova(tmp.single_intercept, tmp.back_elim)$"Pr(>Chisq)"[2]

    if (pval <= 0.05) {
        tmp_final <- tmp.back_elim
    } else {
        tmp_final <- tmp.single_intercept
    }

    model.formula.alldata[[i]] <- formula(tmp_final)

}
model.formula.alldata
fla <- formula(model.formula.alldata[["SelMeth"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "SelMeth", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Significance of random effects terms for SelMeth Rubric")
tmp.single_intercept <- update(tmp, . ~ . +
    1 - as.factor(Rater))
knitr::kable(anova(tmp.single_intercept,
    tmp), caption = "Significance of the Rater intercept term for SelMeth")
tmp.fixed_interactions <- update(tmp, . ~
    . + as.factor(Rater) * Semester - Semester)
knitr::kable(anova(tmp, tmp.fixed_interactions),
    caption = "ANOVA for the interaction terms")
```

```
summary(tmp)
fla <- formula(model.formula.alldata[["CritDes"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "CritDes", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Significance of random effects terms for CritDes Rubric")
tmp.single_intercept <- update(tmp, . ~ . +
    1 - as.factor(Rater))
knitr::kable(anova(tmp.single_intercept,
    tmp), caption = "Significance of the Rater intercept term for CritDes")
summary(tmp)
fla <- formula(model.formula.alldata[["InterpRes"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "InterpRes", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Significance of random effects terms for InterpRes Rubric")
tmp.single_intercept <- update(tmp, . ~ . +
    1 - as.factor(Rater))
knitr::kable(anova(tmp.single_intercept,
    tmp), caption = "Significance of the Rater intercept term for InterpRes")
summary(tmp)
fla <- formula(model.formula.alldata[["VisOrg"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "VisOrg", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Significance of random effects terms for VisOrg Rubric")
tmp.single_intercept <- update(tmp, . ~ . +
    1 - as.factor(Rater))
knitr::kable(anova(tmp.single_intercept,
    tmp), caption = "Significance of the Rater intercept term for VisOrg")
summary(tmp)
fla <- formula(model.formula.alldata[["InitEDA"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "InitEDA", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for InitEDA Rubric")
summary(tmp)
fla <- formula(model.formula.alldata[["RsrchQ"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "RsrchQ", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for RsrchQ Rubric")
summary(tmp)
fla <- formula(model.formula.alldata[["TxtOrg"]])
tmp <- lmer(fla, data = tall.nonmissing[tall.nonmissing$Rubric ==
    "TxtOrg", ])
knitr::kable(round(summary(tmp)$coef, 2),
    caption = "Estimate and Significance of random intercept term for TxtOrg Rubric")
summary(tmp)
# Order RsrchQ, CritDes, InitEDA,
# SelMeth, InterpRes, VisOrg, TxtOrg
row1 <- c(2.35, 0, 2.44, 0, 0, 0, 2.59)
row2 <- c(0, 1.69, 0, 2.25, 2.7, 2.38, 0)
```

```
row3 <- c(0, 2.11, 0, 2.23, 2.59, 2.65, 0)
row4 <- c(0, 1.89, 0, 2.03, 2.14, 2.28, 0)
row5 <- c(0, 0, 0, -0.36, 0, 0, 0)
row6 <- c(0.07, 0.43, 0.37, 0.09, 0.06, 0.29,
    0.09)
row7 <- c(0.28, 0.24, 0.17, 0.11, 0.25, 0.15,
    0.4)
combined <- as.matrix(rbind(row1, row2, row3,
    row4, row5, row6, row7))
rownames(combined) <- c("Intercept", "Rater 1",
    "Rater 2", "Rater 3", "SemesterS19",
    "Variance by Artifact", "Variance of Residuals")
colnames(combined) <- c("RsrchQ", "CritDes",
    "InitEDA", "SelMeth", "InterpRes", "VisOrg",
    "TxtOrg")
knitr::kable(combined, caption = "Coefficients for the Final Models for each Rubric")
p <- summary(lmer(as.numeric(Rating) ~ 1 +
    (0 + Rubric | Artifact) + as.factor(Rater) +
    Semester + Sex + Repeated + Rubric +
    as.factor(Rater) * Semester * Rubric *
        Sex * Repeated, data = tall.nonmissing))
# Interesting interactions:
# (Intercept); Rater 3 and InterpRes;
# Rater 3 and RsrchQ; Rater
# 2,SemesterS19, and TxtOrg; Rater 2
# ,SemesterS19, and VisOrg; Rater 2,
# SexF, and InitEDA; Rater 2, SexF, and
# TxtOrg; Rater 2, SexF, and VisOrg;
# Rater 2, SexF, Repeated, and TxtOrg;
# Rater 2, SexF, Repeated, and VisOrg
knitr::kable(round(p$coefficients[p$coefficients[,
    3] >= 2 | p$coefficients[, 3] <= -2,
    ], 2), caption = "Significant Fixed Effects and Interaction Terms")
# detach(package:plyr)
ratings2 <- ratings[is.na(ratings$CritDes) ==
    FALSE, ]
ratings2 <- ratings2[is.na(ratings2$VisOrg) ==
    FALSE, ]

box <- ratings2 %>%
    mutate(ratingsum = as.numeric(RsrchQ) +
        as.numeric(CritDes) + as.numeric(InitEDA) +
        as.numeric(SelMeth) + as.numeric(InterpRes) +
        as.numeric(VisOrg) + as.numeric(TxtOrg)) %>%
    select(Semester, Sex, ratingsum)
ggplot(box, aes(x = Semester, y = ratingsum)) +
    geom_boxplot() + ggtitle("Distribution of Ratings by Semester") +
    ylab("Total Ratings Score")
ggplot(box, aes(x = Sex, y = ratingsum)) +
    geom_boxplot() + ggtitle("Distribution of Ratings by Gender") +
    ylab("Total Ratings Score")

knitr::kable(ratings2 %>%
```

```
    group_by(Sex) %>%
    mutate(ratingsum = as.numeric(RsrchQ) +
        as.numeric(CritDes) + as.numeric(InitEDA) +
        as.numeric(SelMeth) + as.numeric(InterpRes) +
        as.numeric(VisOrg) + as.numeric(TxtOrg)) %>%
    summarize(genderrating = sum(ratingsum),
        gendermean = mean(ratingsum), count = genderrating/gendermean),
    caption = "Average Rating of Artifacts by Gender")

knitr::kable(ratings2 %>%
    group_by(Semester) %>%
    mutate(ratingsum = as.numeric(RsrchQ) +
        as.numeric(CritDes) + as.numeric(InitEDA) +
        as.numeric(SelMeth) + as.numeric(InterpRes) +
        as.numeric(VisOrg) + as.numeric(TxtOrg)) %>%
    summarize(semesterrating = sum(ratingsum),
        semestermean = mean(ratingsum), count = semesterrating/semestermean),
    caption = "Average Rating of Artifacts by Semester")
```