Identifying Factors in the Rating of Papers

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 implementation of a new "General Education" program for undergraduates and 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 the factors that influence the rating of a paper for each of the 7 rubrics. For some of the rubrics, 3 of them to be exact, there were no external factors that influence the ratings of a paper, but for the other 4 rubrics, various factors and interactions were found necessary in predicting the rating of an artifact. Overall, many conclusions can be made about the ratings of the artifacts, the raters' grading patterns, and the factors involved in predicting an artifacts rating in a particular rubric, all of which will be discussed in later sections of this study.

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. As such, the college has been curious with how ratings are given out to students when they write papers. For this study, past artifacts and their ratings will be analyzed to see if there are any interesting patterns and to answer four questions: If the distribution of ratings are the same across each rubric and rater, if raters tend the same scores for the same rubric and the same artifact, how internal factors such as the contents of the 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.

Data

The data for this study directly comes from 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 appro- priate 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=1 or if the Overlap column has a value present. In total, there will be four datasets used for this study.

Methods

To begin, the first question regarding the distribution of ratings by rubric and by rater only requires some simple Exploratory Data Analysis. A simple table of means and standard deviations for each rubric and rater were made on both the overlapping artifacts and the full dataset to compare the distributions for each rubric and each rater. Then, the frequency of each rating in both the overlapping and full dataset was visualized in a bar chart.

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. The ICC calculates the percentage of agreement between all the raters and how they rate each artifact, this is a value between 0 and 1 where the higher the ICC, the more in agreement the three raters are. Next, the exact percent 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. The result is a matrix and the numbers down the main diagonal will be summed and divided by 13. 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.

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 all possible fixed effects, random effects, and interaction terms. 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.

For the final part of the study, a couple other EDA tables were made to see if there was anything else that could be deemed interesting for this study. In this case, two tables were created to see the average rating between genders and between semesters across all of the rubrics.

Results

Starting with the mean and standard deviation tables in Appendix A and below, 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.

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

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. Looking at the plots from Appendices B,C,D,and E, it appears that most of the rubrics follow a normal distribution centered around a rating of 2, the only exceptions being the Critique Design rubric which had mostly 1's and InterpRes and TxtOrg which had mostly 3's. 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.

In the table below and in Appendix F, a table of ICCs and percent exact agreement can be found to answer the second question.

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 0.2 on both the overlapping artifacts and the full dataset. For the remaining rubrics, the ICCs don't go any higher than 0.7. Between each pair of raters, their percent exact agreements are usually in the range of 0.5 and 0.8 for most of the rubrics, the only main exceptions are between raters 1 and 2 on the Research Question rubric with an agreement of 0.38 and on the Select Methods rubric with an agreement of 0.92.

For the third question, every combination of fixed effects on the random intercept model, considering all rubrics at once, were presented in Appendix G. Each model was shortened to a two letter name (aa,ab,ac...) and tested through ANOVA. Running all fixed effect combinations on only the overlapping artifacts resulted in the best model containing only Repeated as a fixed effect since it had the smallest BIC and none of the Chi-square models were significant. However, since this data is already on

the overlapping artifacts, it's not going to be meaningful in predicting the rating. The next best model is the Rater only model since it has the second lowest BIC and the lowest AIC. However once again, looking at the summary in Appendix H shows that rater is not significant to the model since it has a t-value between -2 and 2. This means that the intercept only model is likely the best model here. Running the same process on the full dataset (models named ba,bb,bc…) in Appendix I gives two possible good fixed effect models, Rater only and Rater with Sex. Running summary on both of these functions display that Rater is significant to the model but not Sex. So as of right now Rater appears to be the only external factor significantly influencing the rating of an artifact regardless of rubric.

We consider each rubric separately in Appendices L and M as shown in the outputs below.

```
$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)
```

Formulas for Fixed Effect Models in Overlapping 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 + Sex + (1 |
    Artifact) - 1

$TxtOrg
as.numeric(Rating) ~ (1 | Artifact)

$VisOrg
as.numeric(Rating) ~ as.factor(Rater) + (1 | Artifact) - 1
```

Formulas for the Fixed Effect Models in All Artifacts

First, using only the overlapping artifacts, none of the fixed effects are significant in predicting rating in all of the rubrics, so all rubrics follow a random intercept model. When using the same method on the full dataset, a few fixed effects come up as significant for about half of the rubrics. For three of the rubrics, Initial EDA, Research Question, and Text Organization, no fixed effects came up as significant so the predictive model for these rubrics are simply the random intercept models. Out of the remaining rubrics, Critique Design, Interpret Results, and Visual Organization have rater as a significant fixed effect. Running the summary and anova tests on these rubrics give extremely high t-values and low p-values for the rater effect validating the significance of the fixed effect. Lastly for the Select Methods rubric, two fixed effects were significant, rater and semester. For this rubric, summary statistics and anova tests were conducted to see if the fixed effects were significant like before, but an additional ANOVA test was carried out to see if an interaction term is needed for the model between rater and semester. The summary and ANOVA tests came out significant for rater and semester separately, but insignificant for the interaction term, as such the final model for predicting an artifacts rating in the Select Method rubric contains two fixed effects, rater and semester, and the random intercept.

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, as seen in Appendix O.

Some other interesting results from the data to answer the fourth question are that the average total rating across all of the rubrics between genders are about 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. These tables can be seen below and in Appendix P.

Table 18: 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 19: 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 identify any patterns between the ratings of 91 artifacts and how various factors such as the rater, sex of the student, and the semester influence those ratings, to give Dietrich College an at Carnegie Mellon Unviersity an insight on their “General Education” program.

From the mean and standard deviation tables at the beginning of the Technical Appendix, it can be quite difficult to deduce any patterns between the raters or the rubrics. The patterns that were mentioned in the results section were not consistent at all and had many exceptions across both rubrics and raters. Overall, this means that the average rating and standard deviations of the ratings do vary between rubrics and by raters, but that no one rater rates higher or lower than the other two on all rubrics, and that no one rubric has a higher or lower rating than the other rubrics between the raters. The explanation for this variance is likely to be small "noises" and deviations in the frequencies of the ratings, which is to be expected.

Breaking down each rubric individually, the normal distribution seen in the bar plots should be an acceptable outcome for the "General Education" program as the school should strive for most of the student population to have a moderate to satisfactory quality of work with only a handful of students either absolutely excelling or falling behind. The main issue comes in the Critique Design rubric as it's the only rubric that is heavily weighted to the left. A possible explanation for this is that freshman statistics students don't have much in-depth knowledge about study designs and are making things up in their artifacts. Another explanation for this is maybe that the raters are confused about how to rate this specific rubric since the rating table is scored based on evidence while the rubric description is asking the students to evaluate a study design which might not fall under the category of "generating evidence".

Between the raters, they all roughly agree on their scores. While there are some notable differences as mentioned in the results section above, it shouldn't be too much of a concern given the distribution of ratings for each rubric and rater. One important thing to consider is that the raters all come from different departments, which in Dietrich College, includes a wide range of departments such as Economics, English, History, Philosophy, and Statistics. The department a rater comes from has already been shown to be an influencing factor on the ratings of an artifact since a rater in one department may look for small details other departments' raters won't look for. It might be better for the raters to be all from the same department and in the department of the class the raters are rating. As an example, artifacts from a statistics class might benefit the best with three raters from the statistics. This is definitely one of the things that should be explored should this study be repeated or continued in the future.

Overall, the rating system made for the "General Education" program is very satisfactory right now, as there is a good amount of consistency when it comes to how ratings are given out. However, there are some glaring potential problems not covered in the data that will need to be further investigated.

References

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

Technical Appexdix

Kevin Yang

11/26/2021

```
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
```

Appendix A: 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

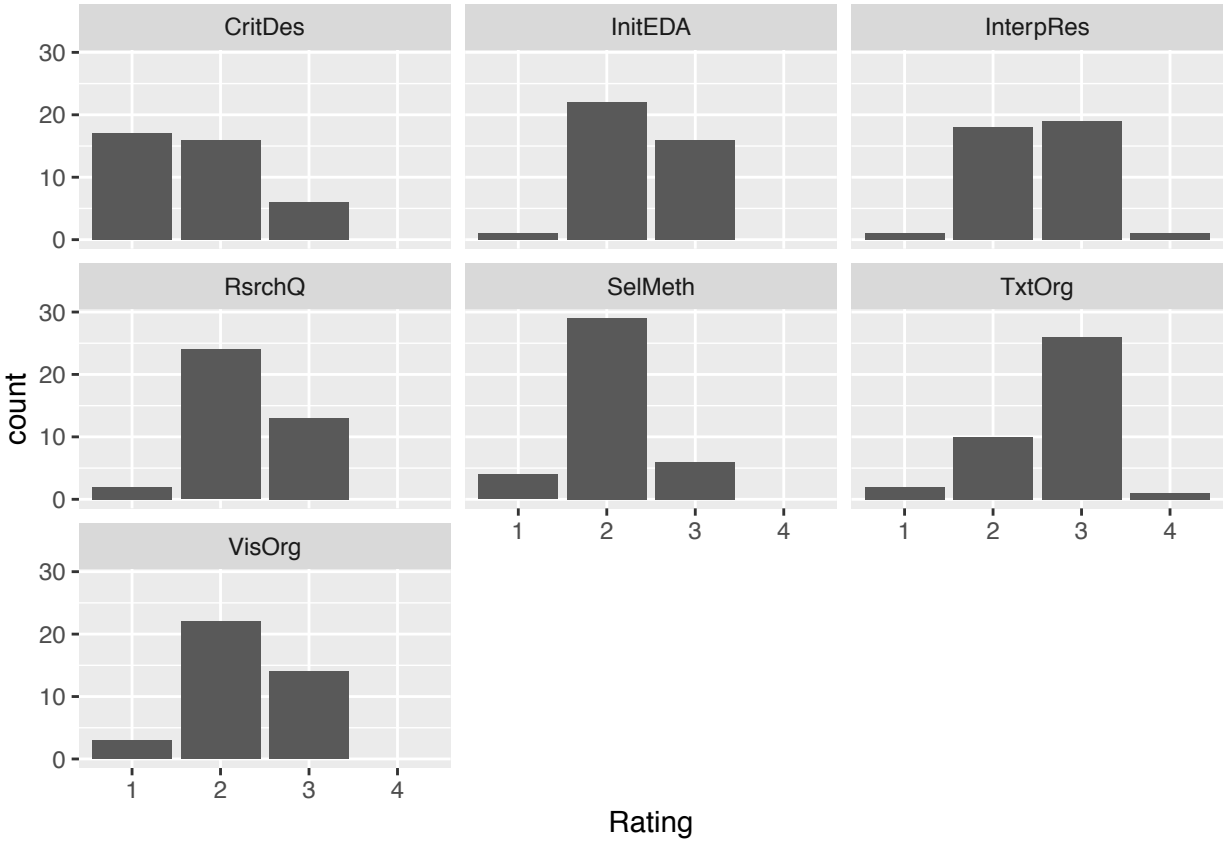
Table 3: Standard Deviation of ratings by Rater in Overlapping Artifacts

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

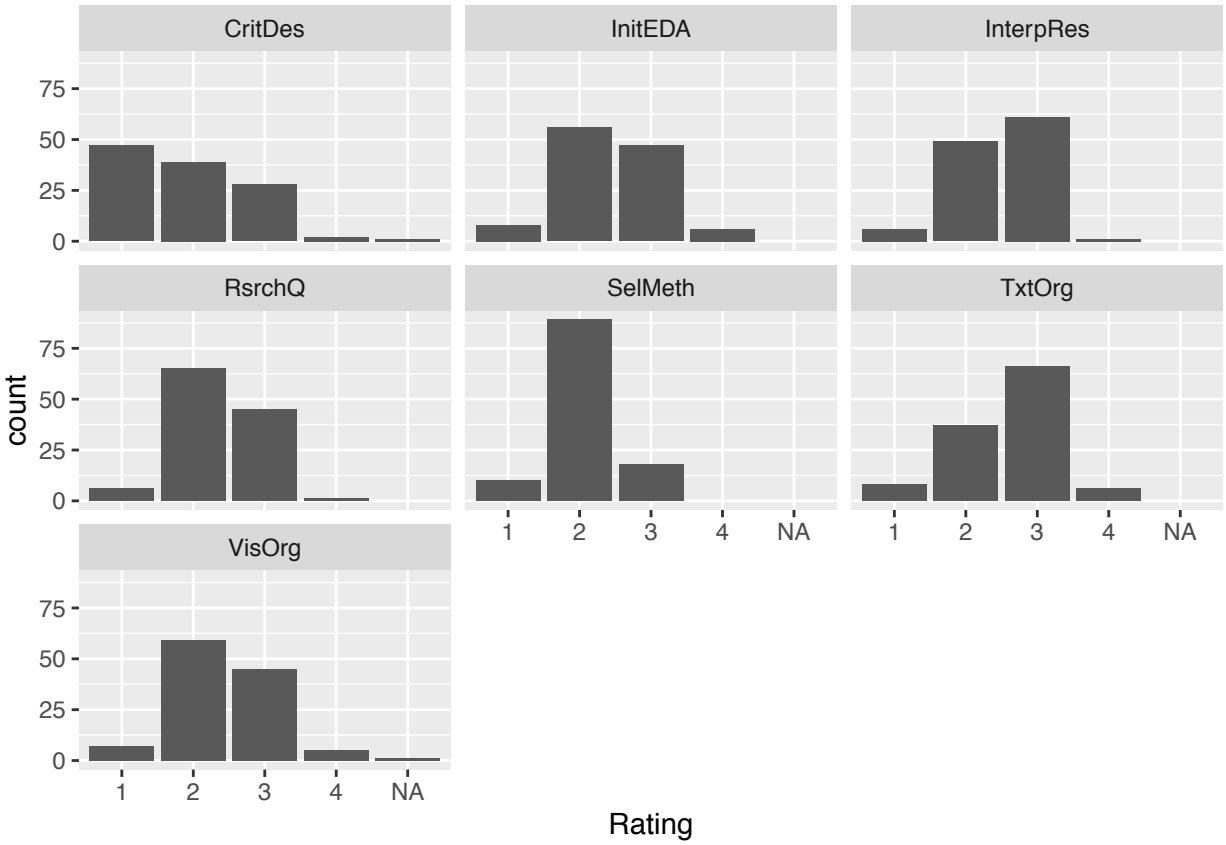
Table 4: Standard Deviaiton of ratings by Rater in full dataset

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

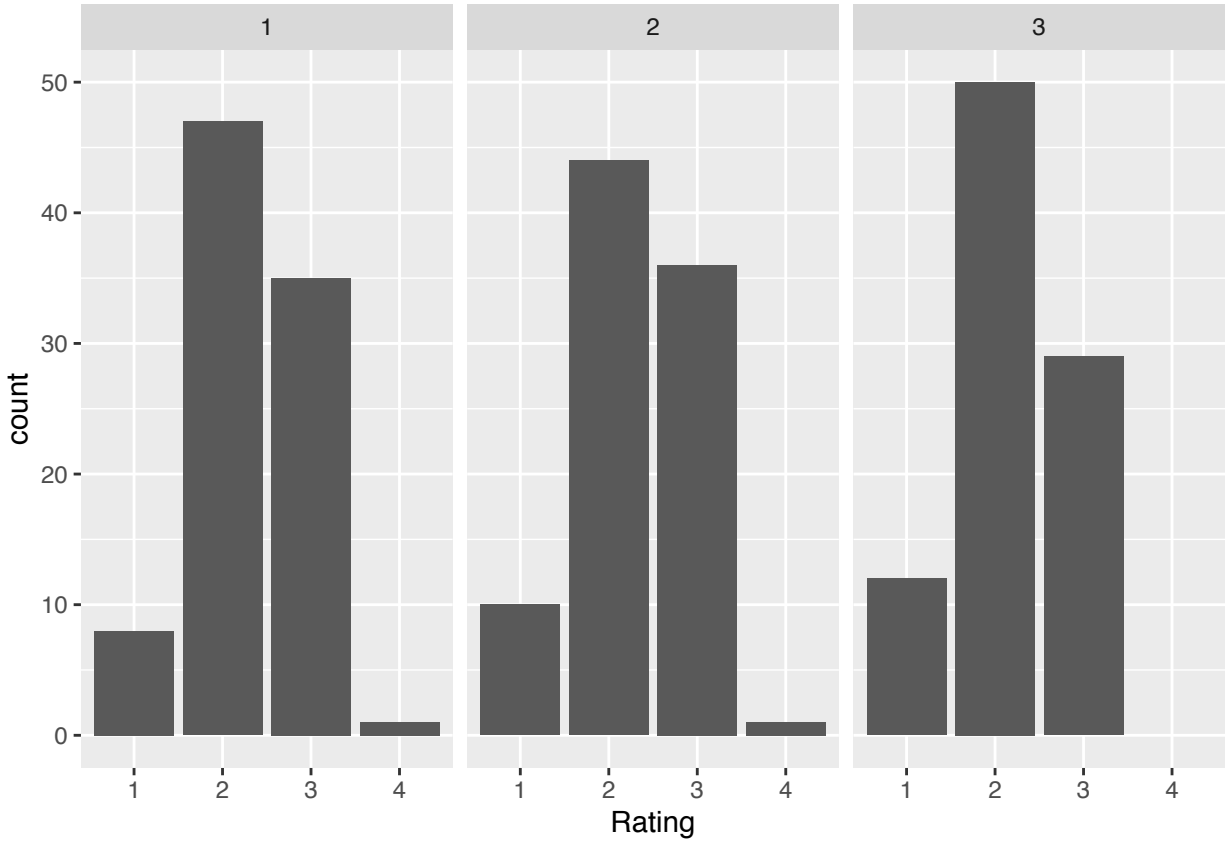
Appendix B: Visual Distribution of Ratings in Overlapping Artifacts



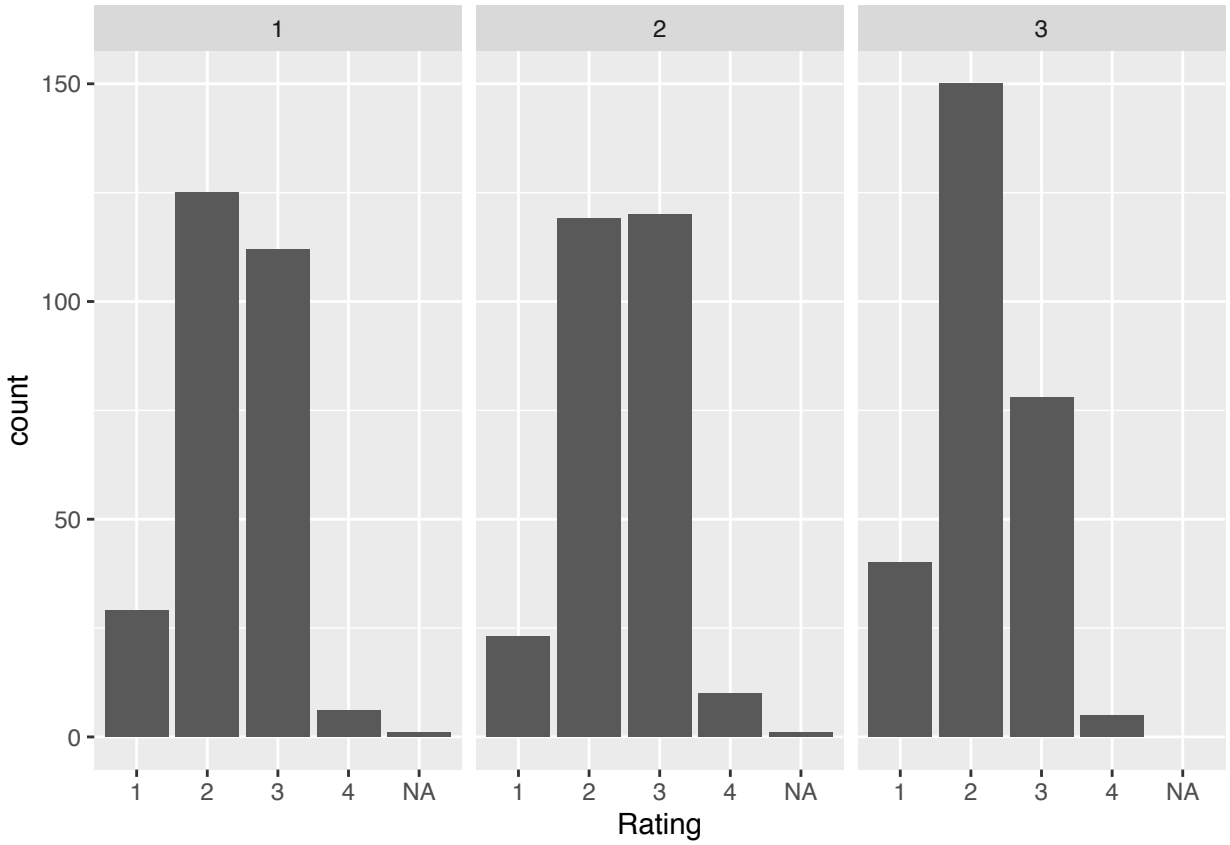
Appendix C: Visual Distribution of Ratings in All Artifacts



Appendix D: Visual distributions of ratings by rater on overlap dataset



Appendix E: Visual distribution of ratings by rater on full dataset



Appendix F: ICC and agreement percentages for each rubric and raters

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

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

Table 6: ANOVA for all rubrics and 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
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

Appendix H: 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
```

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

Table 7: ANOVA for all rubrics and 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

Appendix J: Summary of Rater only Model from Previous Appendix

```
#> 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
```

Appendix K: Summary of Rater and Sex Model from Appendix I

```
#> 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
```

Appendix L: 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)
```

Appendix M: 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
```

Appendix N: Significance of Random Effects/Interaction Terms

SelMeth

Table 8: 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

Table 9: 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

Table 10: 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

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 11: 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

Table 12: 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

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 13: 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

Table 14: 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

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 15: 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

Table 16: 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

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
```

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

Table 17: 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
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

Appendix P: Average Rating of Artifacts based on Gender and Semester

Table 18: 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 19: Average Rating of Artifacts by Semester

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

Code Appendix

```
knitr::opts_chunk$set(comment = "#>", tidy.opts = list(width.cutoff = 40),
    tidy = TRUE)
library(arm)
library(lme4)
library(plyr)
library(tidyverse)
library(performance)
library(LMERConvenienceFunctions)
library(RLRsim)
setwd("~/Documents/College/Semester 9/Applied Linear Modeling/ALM HW10")
ratings <- read.csv("ratings.csv")
tall <- read.csv("tall.csv")  # Rows 5,122,239,356,473,590,707 have NAs
ratingsoverlap <- ratings[is.na(ratings$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")
```

```
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$Repeated == 1, ]
ggplot(talloverlap, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar()
ggplot(tall, aes(x = Rating)) + facet_wrap(~Rubric) +
    geom_bar()
ggplot(talloverlap, aes(x = Rating)) + facet_wrap(~Rater) +
    geom_bar()
ggplot(tall, aes(x = Rating)) + facet_wrap(~Rater) +
    geom_bar()
talloverlap$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$Rubric == "TxtOrg",
    ]
h <- lmer(Rating ~ 1 + (1 | Artifact), data = RsrchQall)
j <- lmer(Rating ~ 1 + (1 | Artifact), data = CritDesall)
k <- lmer(Rating ~ 1 + (1 | Artifact), data = InitEDAall)
l <- lmer(Rating ~ 1 + (1 | Artifact), data = SelMethall)
m <- lmer(Rating ~ 1 + (1 | Artifact), data = InterpResall)
n <- lmer(Rating ~ 1 + (1 | Artifact), data = VisOrgall)
o <- lmer(Rating ~ 1 + (1 | Artifact), data = TxtOrgall)
rubricnames <- c("RsrchQ", "CritDes", "InitEDA",
    "SelMeth", "InterpRes", "VisOrg", "TxtOrg")
icc1 <- rbind(icc(a), icc(b), icc(c), icc(d),
    icc(e), icc(f), icc(g))[, 1]
icc2 <- rbind(icc(h), icc(j), icc(k), icc(l),
    icc(m), icc(n), icc(o))[, 1]
icctable <- cbind(rubricnames, icc1, icc2)
colnames(icctable) <- c("Rubric", "ICC for Overlaps",
    "Icc for Full")
table(RsrchQ[RsrchQ$Rater == 1, 8], RsrchQ[RsrchQ$Rater ==
    2, 8])  # 5 matches
table(RsrchQ[RsrchQ$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 and 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 and 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)
```

```
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, ]
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")
```