Analyzing Variability in Student Generated Statistical Artifacts

Maxine Graves
mgraves@andrew.cmu.edu

Abstract

Carnegie Mellon University (CMU) is interested in gauging the efficacy of general education (GE) courses. In order to address this interest, an analysis of a dataset including ratings on pertinent rubric items of student generated artifacts from a statistics GE was carried out. The analysis was comprised of exploratory data analysis (EDA), graphing, multi-level mixed modelling, and analysis of variance. Results showed that distributions of rubric item ratings are somewhat contingent on rubric item and rater, there is a relatively high level of rating agreement among raters, Rater and Rubric are the two most important variables in determining Rating, and that high rating agreement among raters does not preclude a level of disagreement among the same. Overall, creating a formalized approach to rating would augment one's ability to judge the efficacy of GE courses in imparting specified skills.

Introduction

As is the case at most colleges, Carnegie Mellon University (CMU) places a strong emphasis on general education requirements (GEs). Like the name suggests, these courses are meant to give all students foundational knowledge that will be beneficial regardless of chosen major. Given that GEs are an important facet of university curricula, ensuring that these required courses are able to impart the expected skills is pertinent. One way of determining whether GEs meet expectations is by having university faculty rate student artifacts on rubrics indicative of course efficacy. To this end, this paper seeks to answer the following questions regarding this form of metric:

1. Is the distribution of rubric ratings constant across all rubrics and raters?
2. How much agreeance is there in rating between raters at the rubric level?
3. Do any variables included in the overall dataset seem to be related to ratings and are there any interactions among variables?
4. Regarding the data, what else can be said?

Data

Data used for the present paper comes from CMU's Dietrich College. It includes 15 variables on 91 "artifacts" (statistical papers) written by students. While each artifact received ratings from at least one rater, a subset of 13 artifacts received ratings from all three raters. For a comprehensive list of variables, see Table 1. During the exploratory phase of the paper, three "N/A" values were found, one student did not select "M" of "F" for the Sex variable, another was missing a rating for the CritDes rubric item, and a third was missing a rating for VisOrg. In the first case, a third Sex level, "- -", was created to respect the possibility that this student does not identify as either male or female and therefore decided to choose neither of the two options presented. In the second and third cases, the missing values were replaced with the modes of each rubric rating (both two). See Section A of Technical Appendix for more information.

Variable	Definition
X	Row number
Rater	Identifies rater who rated specific artifact
Sample	Sample number
Overlap	Determines which rater(s) saw which artifacts
Semester	Denotes semester during which artifact was written
Sex	Gender or sex of student who wrote artifact

RsrchQ	Rating on Research Question (rubric item dealing with the generation/critique of a research question)
CritDes	Rating on Critique Design (rubric item dealing with student's ability to critique experimental design of a specified research question)
InitEDA	Rating on Initial EDA (rubric item dealing with the production of exploratory data analysis)
SelMeth	Rating on Selected Methods (rubric item dealing with the appropriate selection of statistical methods for a specified question)
InterpRes	Rating on Interpreting Results (rubric item dealing with results interpretation)
VisOrg	Rating on Visual Organization (rubric item on efficacy of selected visuals in artifact)
TxtOrg	Rating on Text Organization (rubric item on efficacy of written text in artifact)
Artifact	Artifact ID
Repeated	Denotes whether an artifact was seen by one or three raters (0 and 1, respectively)

Table 1 Comprehensive look at variables included in dataset.

Turning next to some exploratory data analysis (EDA), five number summaries were created for each variable in the dataset. As can be seen in Table 2, each rater saw an equal number of artifacts, 26 of which were seen only by that one rater and 13 were seen by all three raters. Further, looking at the breakdown of artifacts by semester, one sees that a large majority of artifacts were produced during the Fall 2019 semester. In addition, turning to the summaries of the seven rubric items, one can see that all seven items have the lowest possible rating as their minimum (a rating of one) and all except SelMeth have a maximum of the highest possible rating (a rating of four). The means and medians of each rubric fall within .5 of each other, which translates to relatively minimal skewing in most rating distributions.

```
       X          Rater       Sample          Overlap      Semester   Sex        RsrchQ          CritDes          InitEDA
 Min.   :  1     1:39    Min.   :  1.00   Min.   : 1    Fall  :83   --: 1   Min.   :1.00   Min.   :1.000   Min.   :1.000
 1st Qu.: 30     2:39    1st Qu.: 31.00   1st Qu.: 4    Spring:34   F :64   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000
 Median : 59     3:39    Median : 60.00   Median : 7                M :52   Median :2.00   Median :2.000   Median :2.000
 Mean   : 59             Mean   : 59.89   Mean   : 7                        Mean   :2.35   Mean   :1.872   Mean   :2.436
 3rd Qu.: 88             3rd Qu.: 89.00   3rd Qu.:10                        3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000
 Max.   :117             Max.   :118.00   Max.   :13                        Max.   :4.00   Max.   :4.000   Max.   :4.000
                                          NA's   :78
    SelMeth         InterpRes          VisOrg          TxtOrg          Artifact      Repeated
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   01      : 3   Min.   :0.0000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   010     : 3   1st Qu.:0.0000
 Median :2.000   Median :3.000   Median :2.00   Median :3.000   011     : 3   Median :0.0000
 Mean   :2.068   Mean   :2.487   Mean   :2.41   Mean   :2.598   012     : 3   Mean   :0.3333
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.:3.000   013     : 3   3rd Qu.:1.0000
 Max.   :3.000   Max.   :4.000   Max.   :4.00   Max.   :4.000   02      : 3   Max.   :1.0000
                                                                (Other):99
```

Table 2 Five number summaries of all variables in dataset.

This point is reinforced by the histograms of ratings on each rubric item for all artifacts that can be found in Figure 1. As can be seen in the figure, all rubric items have relatively normal distributions with the most data falling around two or three, except for CritDes which has a right skewed distribution and with artifacts receiving a 1 on this rubric item more than any other score.

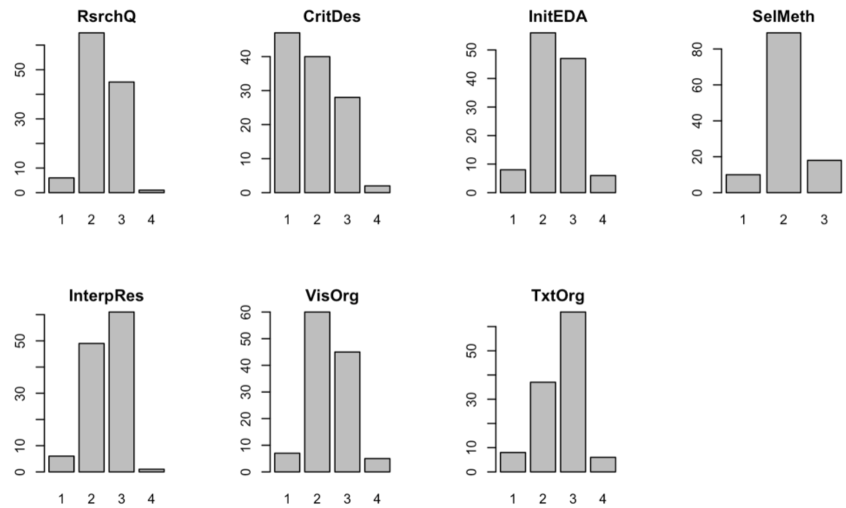


Figure 1 Histograms of ratings broken down by rubric item for all 91 artifacts.

Methods

A variety of statistical methods were employed to answer the aforementioned research questions. Said methods are broken down by question below.

1. Is the distribution of rubric ratings constant across all rubrics and raters?
   Methods used include extensive EDA and graphing.
   variables: Rater and Rubric
2. How much agreeance is there in rating between raters at the rubric level?
   Methods used include intraclass correlation (ICC) and percent exact agreement among raters. Percent exact agreement was calculated by counting the number of times each rater gave the same rating on the same rubric item on the same artifact as another rater, dividing by the total number of artifacts seen by all three raters and taking the sum.
   variables: Rating, Rater, Rubric, and Artifact
3. Do any variables included in the overall dataset seem to be related to ratings and are there any interactions among variables?
   Methods used include multi-level mixed modelling, analysis of variance, and Bayesian Information Criterion (BIC).
   variables: Rating, Rater, Rubric, Artifact, Semester, Sex, and Repeated
4. Regarding the data, what else can be said?
   Methods used include percent disagreement among raters.
   variables: Rating, Rater, Rubric, and Artifact

Results

The following section will be divided into four sections, one for each research question.

Is the distribution of rubric ratings constant across all rubrics and raters?

In order to address the first half of the above research question regarding the distribution of rubric ratings across rubrics, when CritDes is barred, the distribution of ratings is relatively constant (see Figure 1). As previously noted, excepting CritDes, all rubrics appear relatively normally distributed, centering around a rating of two or three. A possible interpretation of this is that raters normalized their scoring habits based on the ratings possible. In other words, since the range of possible ratings moves from one to four, raters could have given most students midland scores with only very good and very poor artifacts in a designated rubric receiving a more extreme score of one or four respectively. Further, the right skewing of CritDes may be attributable to a number of different reasons. It is possible that providing a critique of an experimental design is an especially difficult task that students struggle with or it is possible that professors do not spend enough time helping students develop this skill in comparison with other rubric items rated.

In addition, when data is divided into two groups based on how many raters saw a specific artifact, the same patterns as those seen in Figure 1 hold (see Section 1 of Technical Appendix for more information). This in turn ensures that rubric distributions are similar regardless of whether an artifact was seen by all three raters or only one rater. Overall, barring CritDes, the distribution of rubric ratings is relatively constant across rubric items.

Turning next to the second half of the research questions; whether distributions of rubric ratings are constant across raters, there appears to be less overt similarity to that seen in the answer to the first half of the question. As noted above, 13 of the 91 artifacts received ratings from all three raters. These artifacts are of especial import given that they allow for a comparison of rating distributions between raters on the same material. The similarities and differences between raters on common artifacts can be seen in Figure 2.

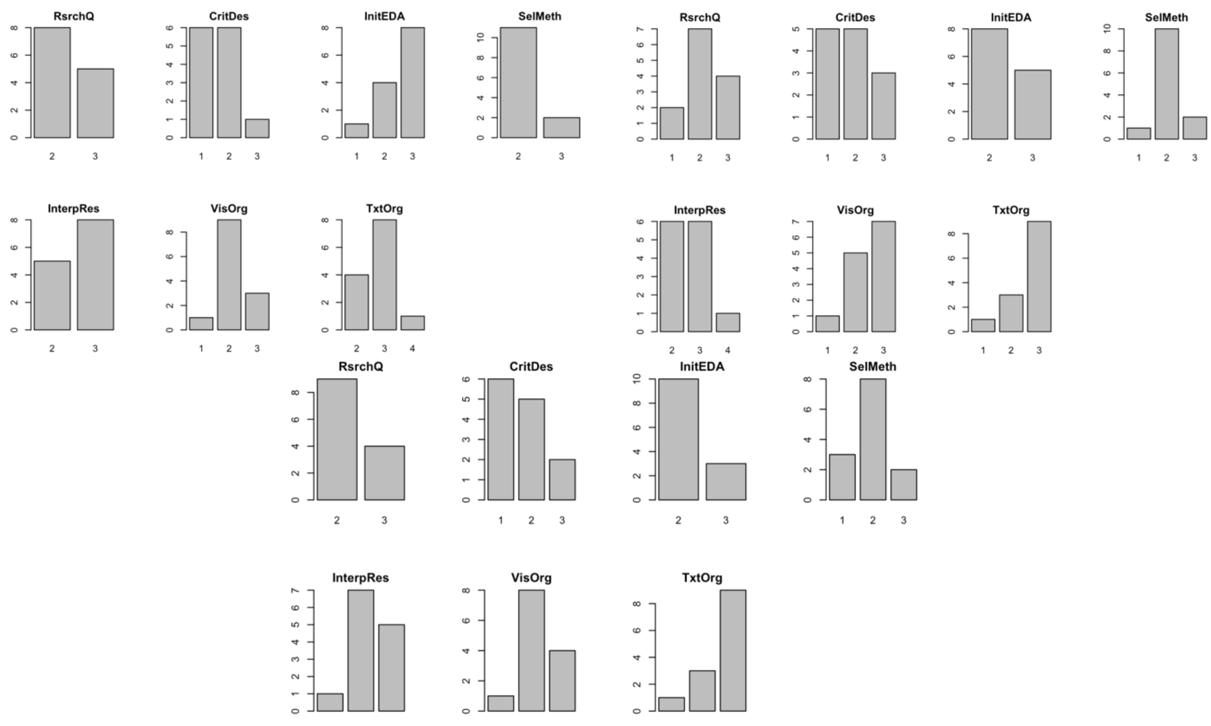


Figure 2 Distribution of ratings given by each rater on the 13 artifacts seen by all raters (from left to right: Rater 1, Rater 2, Rater 3).

While Figure 2 shows that there are notable differences between the rating distributions of each rater on the commonly seen artifacts, it is interesting to note that for all rubric items, at least two of the raters seem to have relatively similar distributions. For example, looking at the ratings distributions for VisOrg, Raters 1 and 3 have very similar distributions, giving most students a 2, a few students a 3, and only one student a 1. Rater 2 on the other hand shows a very left skewed distribution on this rubric, assigning most students a 3, a few students a 2, and one student a 1. Further, the distributions shown in Figure 2 are very similar to the distributions of rubric items when all 91 artifacts are grouped by rater (see Section 1 of the Technical Appendix). The differences in rating distributions across rater may be attributable to the departments which each rater belongs to. It is possible that two professors' fields share similarities regarding some rubric items, with opinions varying more widely on other items.

How much agreeance is there in rating between raters at the rubric level?
By way of answering this question, the ICC and percent exact agreement among raters was generated for each rubric item. Looking at Table 3, one can see that for ratings given to artifacts seen by all three raters, agreement is highest for VisOrg and CritDes. The lowest correlation among these artifacts is for TxtOrg and RsrchQ. In slight contrast, InitEDA shows the highest agreement among raters for all artifacts regardless of how many raters the artifact was seen by, with the other three highest and lowest spots staying the same. It is interesting to note that although its distribution was distinct from the other rubrics, there is still a high level of correlation on ratings for CritDes. Further, while not exactly the same, the ICC values on each rubric item are in the same ballpark for the 13 artifacts seen by all three raters as the correlations for all 91 artifacts. This in turn means that, while there are some differences, generally speaking ICC is relatively stable.

	all_raters	**all_artifacts**
rq_icc	0.1891918	0.2096164
cd_icc	0.5725134	0.6698643
ieda_icc	0.4930784	0.6867310
sm_icc	0.5212845	0.4718910
ir_icc	0.2295821	0.2200241
vo_icc	0.5924748	0.6585680
to_icc	0.1428682	0.1879831

Table 3 Intraclass correlation (ICC) values (left: artifacts rated by all three raters, right: all artifacts regardless of how many raters rated the artifact).

In order to determine the rating similarities between two specific raters, percent exact agreement was calculated for each pair of raters (i.e. exact agreement between raters 1 and 2, 1 and 3, etc.). Looking at Table 4, agreement on most rubric items between all three pairs of raters falls roughly between 53% and 93%. Only one exact agreement percentage falls out of this range, RsrchQ agreement between raters 1 and 2, with an agreement percentage of 38.46%. Put another way, this means that all raters agree with each other at least half of the time on all rubric items, except for raters 1 and 2 on RsrchQ. It is possible that this discrepancy in agreement over research question generation may stem from the department from which raters 1 and 2 come from. The type of research questions asked in a rater's field can vary significantly. Further, while percent exact agreement is an easily interpretable metric for determining agreement among raters, it is incapable of accounting for agreement that is not exact. For instance, returning to the low RsrchQ agreement between raters 1 and 2, it is possible that these raters gave many scores that were similar (for instance 2 and 3, 3and 4, etc.), but that were not identical. Percent exact agreement cannot account for these forms of close agreement.

rubric	raters_12	raters_13	raters_23
rq	0.3846154	0.7692308	0.5384615
cd	0.5384615	0.6153846	0.6923077
ieda	0.6923077	0.5384615	0.8461538
sm	0.9230769	0.6153846	0.6923077
ir	0.6153846	0.5384615	0.6153846
vo	0.5384615	0.7692308	0.7692308
to	0.6923077	0.6153846	0.5384615

Table 4 Percent exact agreement among raters on the 13 artifacts seen by all three raters.

<u>Do any variables included in the overall dataset seem to be related to ratings and are there any interactions among variables?</u>

To answer the above question, multiple mixed level models were fit, taking Artifact as the sole grouping variable. Artifact was deemed an appropriate grouping variable given that artifacts were selected randomly and are therefore considered a representative sample of the total population of artifacts. While models fit ranged considerably in complexity, the final model selected that was deemed best at explaining the relationship between Rating and possible predictors can be seen below.

$$Rating = Rater + Rubric + (0 + Rater\ |Artifact) + (0 + Rubric\ |\ Artifact)$$

In words, the above model includes Rater and Rubric as fixed effects and Rubric and Rater as random effects of Artifact as explanatory variables to predict Rating. Table 5 includes a more in-depth summary of the final model's fixed effects.

	Estimate	Std. Error	t value
(Intercept)	1.9629783	0.0927901	21.1550312
Rater2	0.0069414	0.0783864	0.0885534
Rater3	-0.1608235	0.0676811	-2.3761947
RubricInitEDA	0.5353223	0.0941189	5.6877217
RubricInterpRes	0.5746624	0.0993261	5.7856135
RubricRsrchQ	0.4488698	0.0853845	5.2570407
RubricSelMeth	0.1508231	0.0915866	1.6467810
RubricTxtOrg	0.6748400	0.0985931	6.8446943
RubricVisOrg	0.5210053	0.0978035	5.3270617

Table 5 Coefficients of the fixed effects of the chosen model.

The final model was chosen by first creating three models with varying fixed effects, interactions, and Rubric as a random effect of Artifact. These models were compared based on their BIC values. BIC was chosen as the main metric of comparison between models since the purpose of creating the model is not to predict Rating from the other variables, but rather to determine possible relationships between the variables in the dataset. As this was the aim, interpretability was deemed an important consideration and BIC prefers more interpretable models. The final step of model building compared the existing model with a model including an additional effect. Once more BIC was used to select the final model, which included the additional effect.

	Beta	min(alpha)	max(alpha)
cd1	1.96297829962909	0.964090709463024	3.35318653481723
cd2	3.93289798735599	2.84191804151764	5.63654001488785
cd3	3.7651331191995	2.75171839371924	5.38339875879339
ieda1	4.46127887235246	3.46372367547626	5.59812337589804
ieda2	6.43119856007936	5.32232366754462	7.68588959967795
ieda3	6.26343369192287	4.96913361137758	7.58259576705219
ir1	4.50061895114662	3.91854527060323	4.93243797911713
ir2	6.47053863887352	5.52909246246343	7.14759046475785
ir3	6.30277377071703	5.13831704193551	6.94076493093936
rq1	4.37482641217448	3.58052146777698	5.28271271531842
rq2	6.34474609990138	5.44200256407742	7.09587328872512
rq3	6.17698123174489	5.12591726351563	7.09929590648849
sm1	4.07677968780044	3.75115602618428	4.4227085286392
sm2	6.04669937552733	4.9852566706079	6.85672945391781
sm3	5.87893450737085	5.08162855153722	6.50156725247223
to1	4.60079656491439	3.5376863135765	5.48041389551835
to2	6.57071625264128	5.37104511920654	7.59089524794113
to3	6.4029513844848	5.10411123583649	7.52809595642987
vo1	4.44696189039077	3.4246549554829	5.40435860628462
vo2	6.41688157811767	5.18656603471481	7.53780792508445
vo3	6.24911670996118	4.88457920254964	7.28466666898999

Table 6 Coefficients for betas and corresponding maximum and minimum alpha values for random effects.

FINISH SECTION

Regarding the data, what else can be said?
Since this was a relatively open-ended question, it was decided that it would be interesting to investigate the second research question (i.e. How much agreeance is there in rating between raters at the rubric level?) from a different perspective. As was noted at the end of the results section on research question two, the way in which agreement between specific pairs of raters was calculated excluded the possibility of close, but not exact, rater agreement. By way of addressing this, a form of percent disagreement was fit. Table 7 shows the percentage of ratings by rubric for which two raters scores differed by two or more points (e.g. one rater giving an artifact a score of 3 on a specific rubric item and another giving the same artifact a score of 1 on the same item). The hope was that looking at percent disagreement could act as an addendum to the analysis done earlier for research question two.

While Table 7 denotes a 7.69% rate of disagreement on RsrchQ between raters 1 and 2 (the rubric and rater pairing with the lowest exact agreement, 38.46%), it shows the same rate of disagreement on InterpRes between raters 2 and 3, which received 61.54% exact agreement. What this means is that it is possible for ratings on the same artifact and rubric item from two different raters to concurrently have high exact agreement and a level of disagreement. However, the preceding statement is tempered by the fact that a 7.69% rate of disagreement only amounts to a pair of raters disagreeing on one of the thirteen artifacts seen by both raters.

rubric	raters_12	raters_13	raters_23
rq	0.0769231	0.0000000	0.0000000
cd	0.0769231	0.0000000	0.0000000
ieda	0.0000000	0.0000000	0.0000000
sm	0.0000000	0.0000000	0.0000000
ir	0.0769231	0.0000000	0.0769231
vo	0.0000000	0.0000000	0.0000000
to	0.0769231	0.0769231	0.0000000

Table 7 Percent disagreement among raters on the 13 artifacts seen by all three raters.

Discussion

In broad strokes, the most important takeaways from the present research are that formalizing the way in which raters rate artifacts would lead to the possibility of richer analysis and that based on current data, the most important variables in determining expected rating are Rater and Rubric. Looking at the first takeaway, currently, there is no concrete way to determine whether ratings given by different raters follow the same set of rating conventions. Introducing a form of rating standardization would allow for more fruitful analyses of how successful GEs are in imparting the expected skills. By minimizing unwanted rater variability, the differences seen across rubric items and artifacts could be more defensibly attributed to student success in acquiring specified skills. Minimizing unwanted rater variability may in turn impact the second takeaway. If the way in which raters give ratings is formalized, this may lead to Rater no longer being an important variable in determining rubric Rating, leading to hopefully more insightful conclusions regarding the efficacy of GEs.

Furthermore, it is necessary to discuss the main general limitation of the present research. As previously noted in the methods section, the mode was used to populate rubric ratings for artifacts with N/A values in one of the seven rubric items. While there were only two artifacts with one rating missing each, it is possible that the use of the mode to populate these missing values may have skewed results slightly. Future research may approach the problem of missing rubric ratings differently, perhaps by completely removing artifacts with missing ratings or by having uneven numbers of ratings across rubric items.

Continuing on, the next four sections approach both main takeaways and limitations at the research question level.

<u>Is the distribution of rubric ratings constant across all rubrics and raters?</u>

Generally speaking, distribution of rubric ratings share similarities across all rubrics and raters. Looking more specifically at distributions across rubrics, a high degree of constancy can be seen, with all rubrics except CritDes having a relatively normal distribution centering around 2 or 3. As noted in the Results section, these distributions may be attributable to raters attempting to normalize their scoring practices, giving most students more average scores of 2 or 3 and only very high performing or very low performing artifacts (on a single rubric item) scores of 4 and 1 respectively. Turning next to distributions across raters, again one sees similarities across rubric ratings, however not the same level of consistency. While there are often at least two raters with similar distributions on a given rubric item, this implies there is one rater with a different distribution. These similarities and differences among raters may be due to differences in the import placed on a specific rubric item by the different departments and fields of study raters pertain to.

What the aforementioned seems to necessitate is a higher level of understanding regarding the way in which raters where giving ratings and what impacts these ratings. If it is true that they were attempting to normalize their ratings, it would be worthwhile for CMU to determine whether this was the way they intended raters to score rubric items. On the one hand, normalizing scores may be a good way of ensuring grading consistency and addressing rating extremes. On the other, normalizing scores by nature pushes more artifacts to more central ratings, possibly making raters round up borderline poor artifacts (on a given rubric item) and round down borderline good artifacts (on a given rubric item). Given that the purpose of the data was to determine whether students were successful in developing skills judged by the seven rubrics, it seems counterintuitive to normalize scores since it seems CMU would want as many students to be successful as possible. Further, determining whether department and field of study has an impact on a raters' ratings may lead to a better understanding of ratings distributions across raters. Depending on the emphasis a rater's field of study or department places on a given rubric item, this may have an impact on the way that a rater would rate this item.

<u>How much agreeance is there in rating between raters at the rubric level?</u>
Overall, it seems that depending on metric used, rater agreement can vary. Where ICC values for rubric items for artifacts seen by all raters range from roughly .14 to .60, percent exact agreement is above 50% for all rubric items and rater pairs except RsrchQ agreement between raters 1 and 2 (with 38.46% agreement). Since percent exact agreement is a more informative metric than ICC, it seems safe to say that rater agreement is relatively high on rubric ratings for artifacts seen by all three raters. While it is not possible to extend the above logic to all artifacts, regardless of how many raters saw them, it is concluded that agreement among raters on all 91 artifacts has a high range of correlation dependent on rubric item (around .18 to .69). Again, a possible reason for the discrepancy in ratings may be attributable to the department and field of study a rater is a part of.

Further, it is important to note that percent exact agreement may not be the best metric possible for gauging rater agreement. Percent exact agreement does not account for almost exact agreement, in which raters gave the same rubric item on the same artifact similar scores (for instance a rating of 2 and 3, or 1 and 2). Although not exactly the same rating, score pairings that differ by only one point main point to a relatively high level of agreement.

<u>Do any variables included in the overall dataset seem to be related to ratings and are there any interactions among variables?</u>
As noted in the results section, to answer this question, a multi-level mixed model was fit. The selected model predicting Rating had fixed effects for Rater and Rubric and random effects for the same. In turn, the final model does not deem the variables Repeated, Semester, or Sex as important in determining Rating. This can be interpreted as meaning that the ratings given by raters are not related to how many raters saw an artifact, which semester an artifact was created during, or the sex or gender of the student who created the artifact. Of models tested, Rating is best explained solely by Rater, Rubric, and Artifact.

Again, as previously mentioned, the final model was chosen using BIC. While BIC is known to produce models that are easier to interpret, higher interpretability comes at the cost of lower predictability. For the purposes of this paper, it seemed that CMU would be more interested in having deeper insights about the most important relationships between Rating and other variables, as opposed to being able to predict Rating more accurately from other variables. Had a different metric like Akaike's Information Criterion (AIC) been used, it is possible that a different final model may have been selected. Further, models tested were user-generated; meaning that all models possible to explain Rating were not fit, only a very small subsection of these were fit.

<u>Regarding the data, what else can be said?</u>
FINISH SECTION

References

-how to reference data?

36617 Project 2

Maxine Graves

11/22/2021

sources:
1. https://www.rstudio.com/resources/cheatsheets/
2. https://community.rstudio.com/t/wont-let-me-install-spam-package/90956
-the above blog post was used to help install the LMERConvenienveFunctions package

SECTION A

```
library(arm)
```

```
## Loading required package: MASS

## Loading required package: Matrix

## Loading required package: lme4

##
## arm (Version 1.11-2, built: 2020-7-27)

## Working directory is /Users/maxine/Documents/MSP_Fall_2021/36617
```

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

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:plyr':
##
##     is.discrete, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(LMERConvenienceFunctions) #source 2
library(RLRsim)
library(ggplot2)
library(knitr)
```

2. source 1

```
ratings = read.csv("ratings.csv", header=TRUE)

#checking for NA values
ratings[which(ratings$Sex=="--"), ] #"--" is third level
```

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

```
which(is.na(ratings$Rater)==TRUE)
```

```
## integer(0)
```

```
which(is.na(ratings$Semester)==TRUE)
```

```
## integer(0)
```

```
which(is.na(ratings$Artifact)==TRUE)
```

```
## integer(0)
```

```
which(is.na(ratings$Repeated)==TRUE)
```

```
## integer(0)
```

```
rubrics = ratings[ , c(7:13)]
which(is.na(rubrics)==TRUE)
```

```
## [1] 161 684
```

```
#2 rows 44, and 99 have missing data in rubric variables
ratings[c(44, 99), ]
```

```
##     X Rater Sample Overlap Semester Sex RsrchQ CritDes InitEDA SelMeth
## 44 44     2     45      NA   Spring   F      2      NA       2       2
## 99 99     1    100      NA     Fall   F      2       3       2       3
```

```
##    InterpRes VisOrg TxtOrg Artifact Repeated
## 44         2      2      3       45        0
## 99         3     NA      2      100        0
```

```
rater_2 = ratings %>%
  filter(Rater==2)
table(rater_2$CritDes) #mode is 2
```

```
##
##  1  2  3  4
## 11 13 12  2
```

```
rater_1 = ratings %>%
  filter(Rater==1)
table(rater_1$VisOrg) #mode is 2
```

```
##
##  1  2  3  4
##  1 23 12  2
```

```
ratings[44, ]$CritDes = 2 #set NA value to mode
ratings[99, ]$VisOrg = 2

tall_ratings = read.csv("tall.csv", header=TRUE) %>%
  mutate(Sex = as.character(Sex))
tall_ratings[which(tall_ratings$Artifact=="5"), ]$Sex = "--"
tall_ratings = tall_ratings %>%
  mutate(Sex = as.factor(Sex))
#rows 161 and 684 have missing data
tall_ratings[161, ]$Rating = 2
tall_ratings[684, ]$Rating = 2

#converting Rater to a factor in both datasets
ratings$Rater = as.factor(ratings$Rater)
tall_ratings$Rater = as.factor(tall_ratings$Rater)
```

######################## SECTION 1 ########################
#source 1

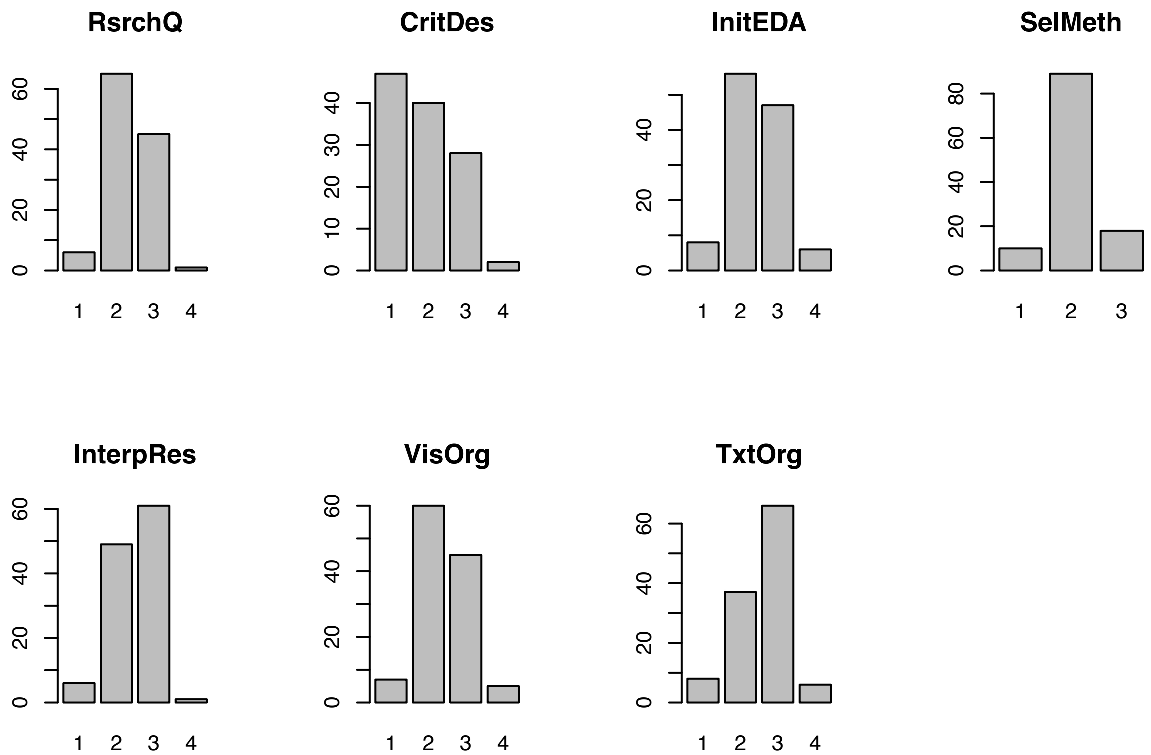
```
summary(ratings)
```

```
##        X         Rater      Sample          Overlap      Semester   Sex
##  Min.   :  1   1:39   Min.   :  1.00   Min.   : 1   Fall   :83   --: 1
##  1st Qu.: 30   2:39   1st Qu.: 31.00   1st Qu.: 4   Spring:34   F :64
##  Median : 59   3:39   Median : 60.00   Median : 7               M :52
##  Mean   : 59          Mean   : 59.89   Mean   : 7
##  3rd Qu.: 88          3rd Qu.: 89.00   3rd Qu.:10
##  Max.   :117          Max.   :118.00   Max.   :13
##                                        NA's   :78
##      RsrchQ         CritDes         InitEDA         SelMeth        InterpRes
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.00   Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.35   Mean   :1.872   Mean   :2.436   Mean   :2.068   Mean   :2.487
##  3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :4.000
##
```

```
##      VisOrg          TxtOrg          Artifact      Repeated
##  Min.   :1.00   Min.   :1.000   O1      : 3   Min.   :0.0000
##  1st Qu.:2.00   1st Qu.:2.000   O10     : 3   1st Qu.:0.0000
##  Median :2.00   Median :3.000   O11     : 3   Median :0.0000
##  Mean   :2.41   Mean   :2.598   O12     : 3   Mean   :0.3333
##  3rd Qu.:3.00   3rd Qu.:3.000   O13     : 3   3rd Qu.:1.0000
##  Max.   :4.00   Max.   :4.000   O2      : 3   Max.   :1.0000
##                                 (Other):99
```

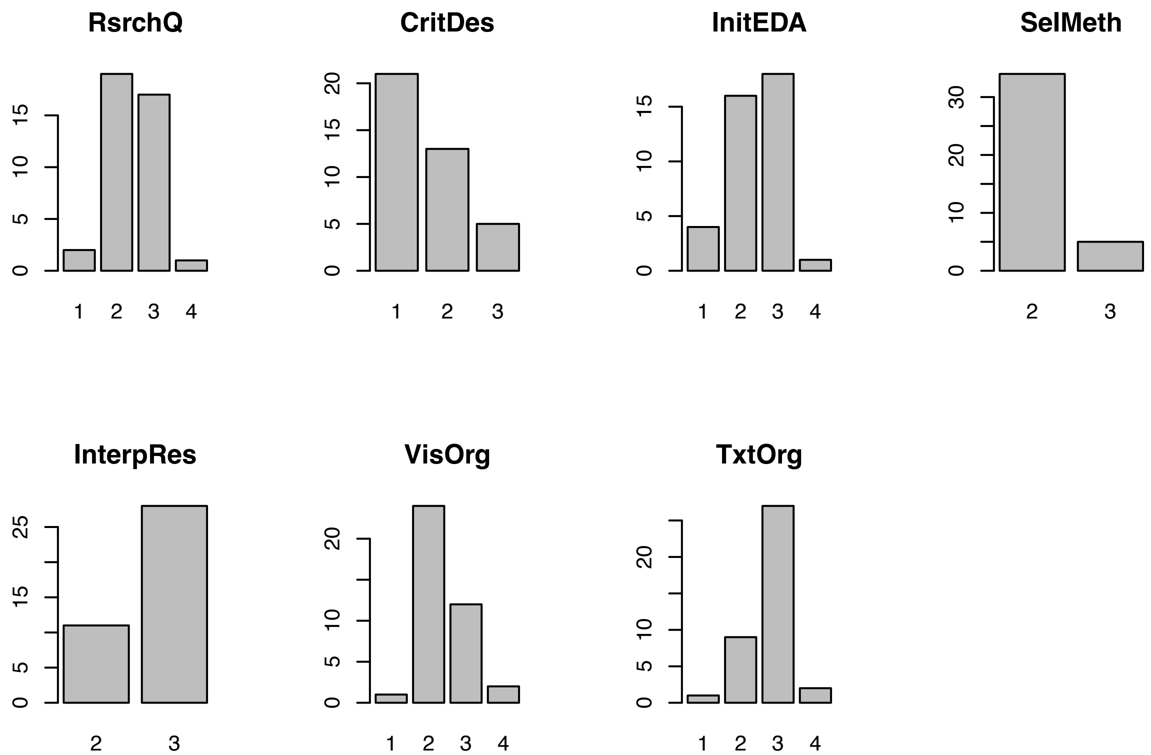
```
#All artifacts grouped by rubric
rubrics = ratings[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
plot(rubrics$RsrchQ,
     main="RsrchQ")
plot(rubrics$CritDes,
     main="CritDes")
plot(rubrics$InitEDA,
     main="InitEDA")
plot(rubrics$SelMeth,
     main="SelMeth")
plot(rubrics$InterpRes,
     main="InterpRes")
plot(rubrics$VisOrg,
     main="VisOrg")
plot(rubrics$TxtOrg,
     main="TxtOrg")

#All artifacts grouped by rubric and rater
rater_1 = ratings %>%
  filter(Rater==1)
rater_1_rubrics = rater_1[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



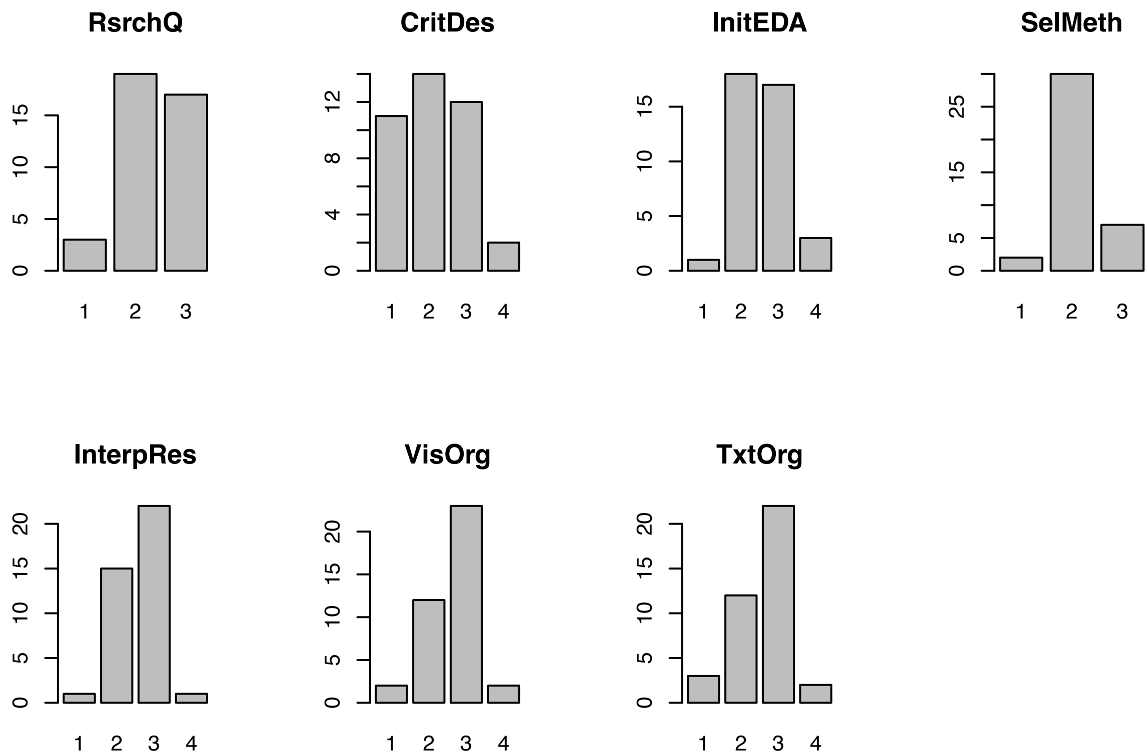

```
plot(rater_1_rubrics$RsrchQ,
     main="RsrchQ")
plot(rater_1_rubrics$CritDes,
     main="CritDes")
plot(rater_1_rubrics$InitEDA,
     main="InitEDA")
plot(rater_1_rubrics$SelMeth,
     main="SelMeth")
plot(rater_1_rubrics$InterpRes,
     main="InterpRes")
plot(rater_1_rubrics$VisOrg,
     main="VisOrg")
plot(rater_1_rubrics$TxtOrg,
     main="TxtOrg")

rater_2 = ratings %>%
  filter(Rater==2)
rater_2_rubrics = rater_2[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



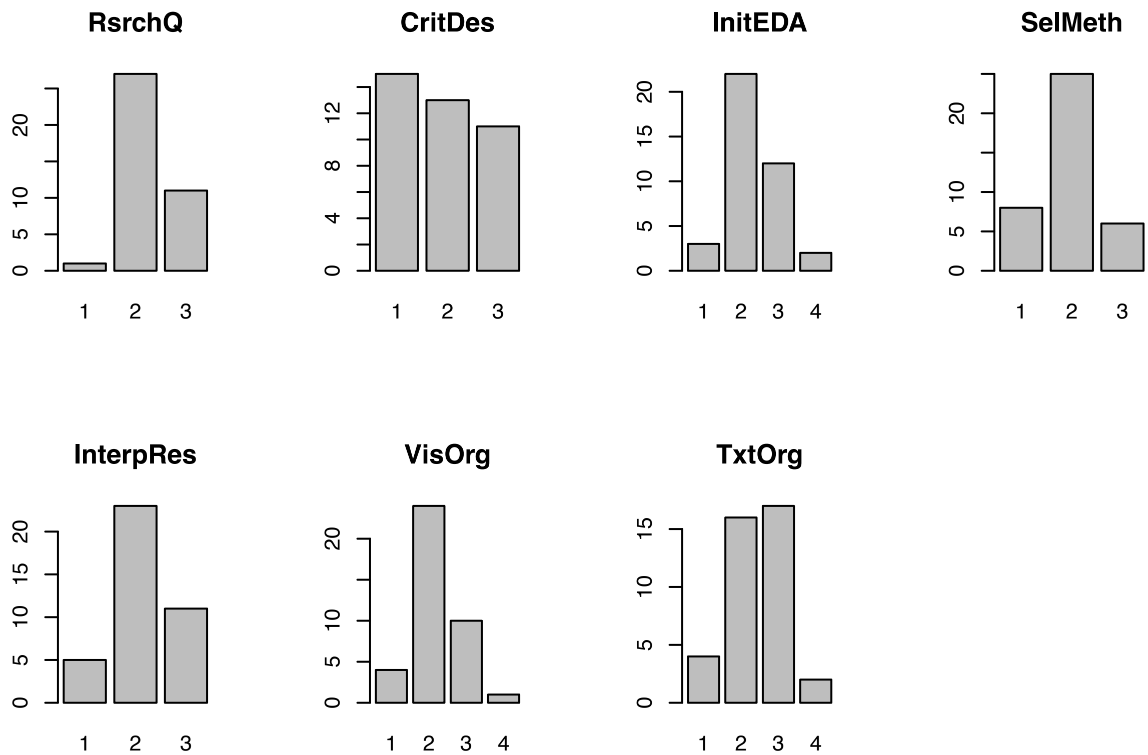

```
plot(rater_2_rubrics$RsrchQ,
     main="RsrchQ")
plot(rater_2_rubrics$CritDes,
     main="CritDes")
plot(rater_2_rubrics$InitEDA,
     main="InitEDA")
plot(rater_2_rubrics$SelMeth,
     main="SelMeth")
plot(rater_2_rubrics$InterpRes,
     main="InterpRes")
plot(rater_2_rubrics$VisOrg,
     main="VisOrg")
plot(rater_2_rubrics$TxtOrg,
     main="TxtOrg")


rater_3 = ratings %>%
  filter(Rater==3)
rater_3_rubrics = rater_3[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



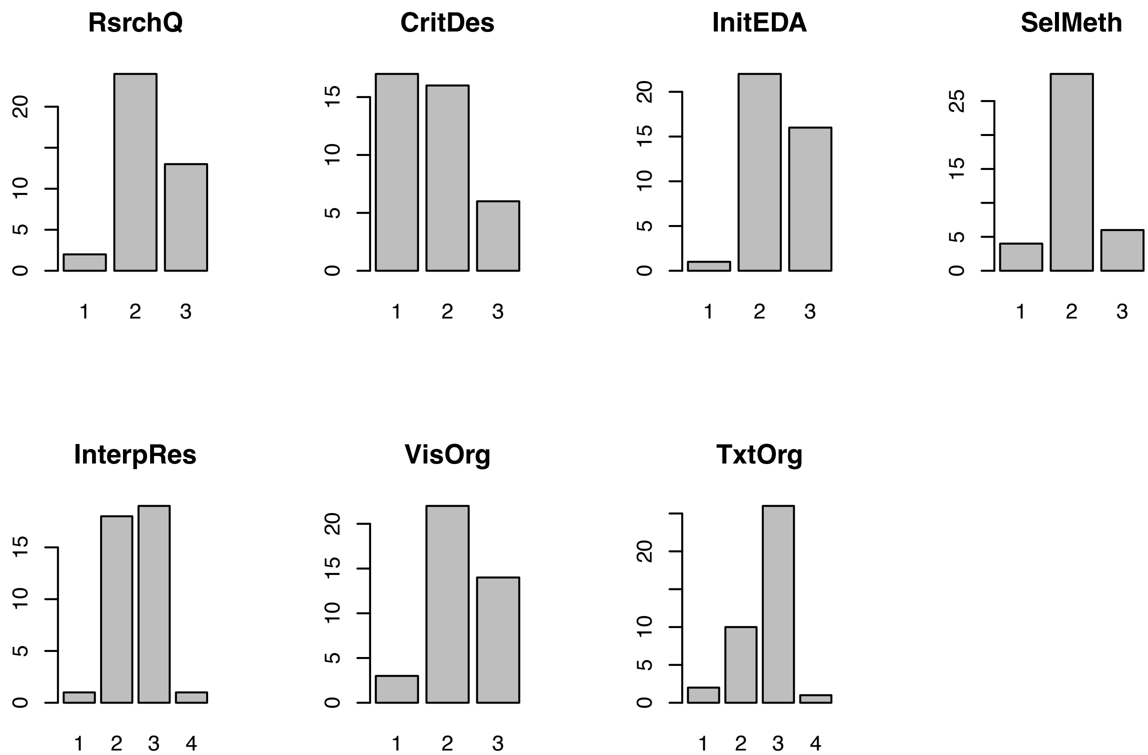

```r
plot(rater_3_rubrics$RsrchQ,
     main="RsrchQ")
plot(rater_3_rubrics$CritDes,
     main="CritDes")
plot(rater_3_rubrics$InitEDA,
     main="InitEDA")
plot(rater_3_rubrics$SelMeth,
     main="SelMeth")
plot(rater_3_rubrics$InterpRes,
     main="InterpRes")
plot(rater_3_rubrics$VisOrg,
     main="VisOrg")
plot(rater_3_rubrics$TxtOrg,
     main="TxtOrg")


#Artifacts seen by all raters grouped by rubric
all_raters_rubrics = ratings %>%
  filter(Repeated==1) %>%
  .[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



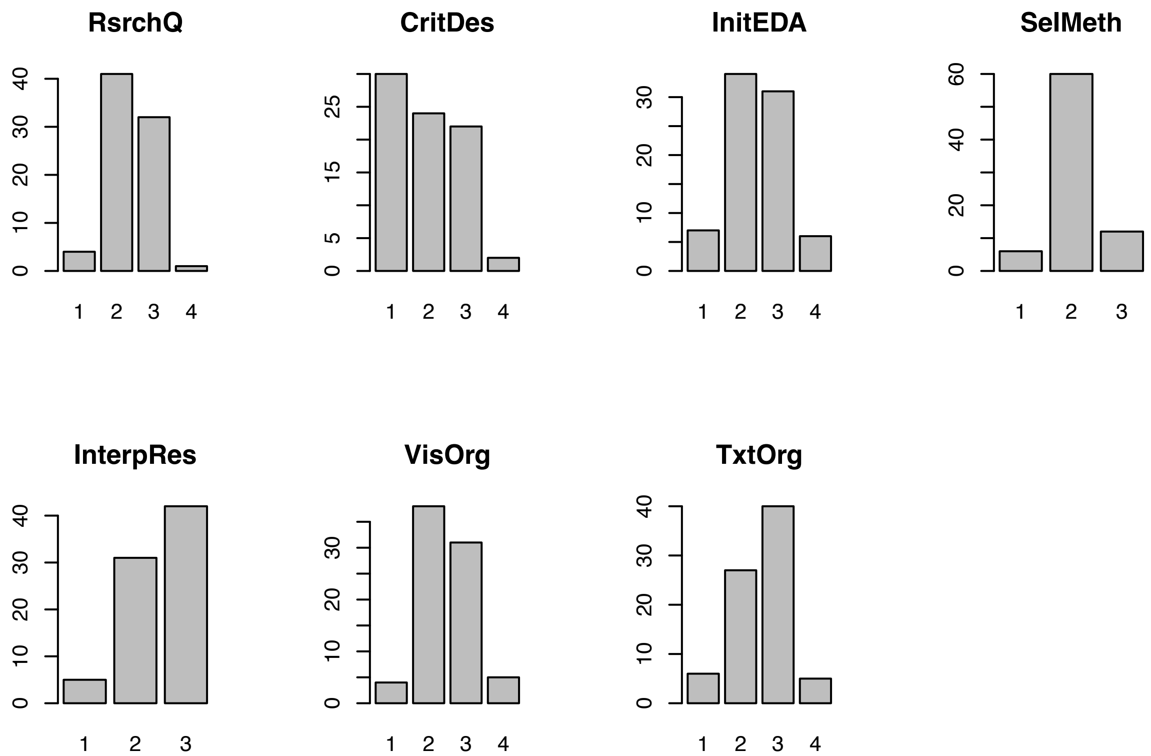

```
plot(all_raters_rubrics$RsrchQ,
     main="RsrchQ")
plot(all_raters_rubrics$CritDes,
     main="CritDes")
plot(all_raters_rubrics$InitEDA,
     main="InitEDA")
plot(all_raters_rubrics$SelMeth,
     main="SelMeth")
plot(all_raters_rubrics$InterpRes,
     main="InterpRes")
plot(all_raters_rubrics$VisOrg,
     main="VisOrg")
plot(all_raters_rubrics$TxtOrg,
     main="TxtOrg")

#Artifacts seen by only one rater grouped by rubric
all_raters_rubrics = ratings %>%
  filter(Repeated==0) %>%
  .[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



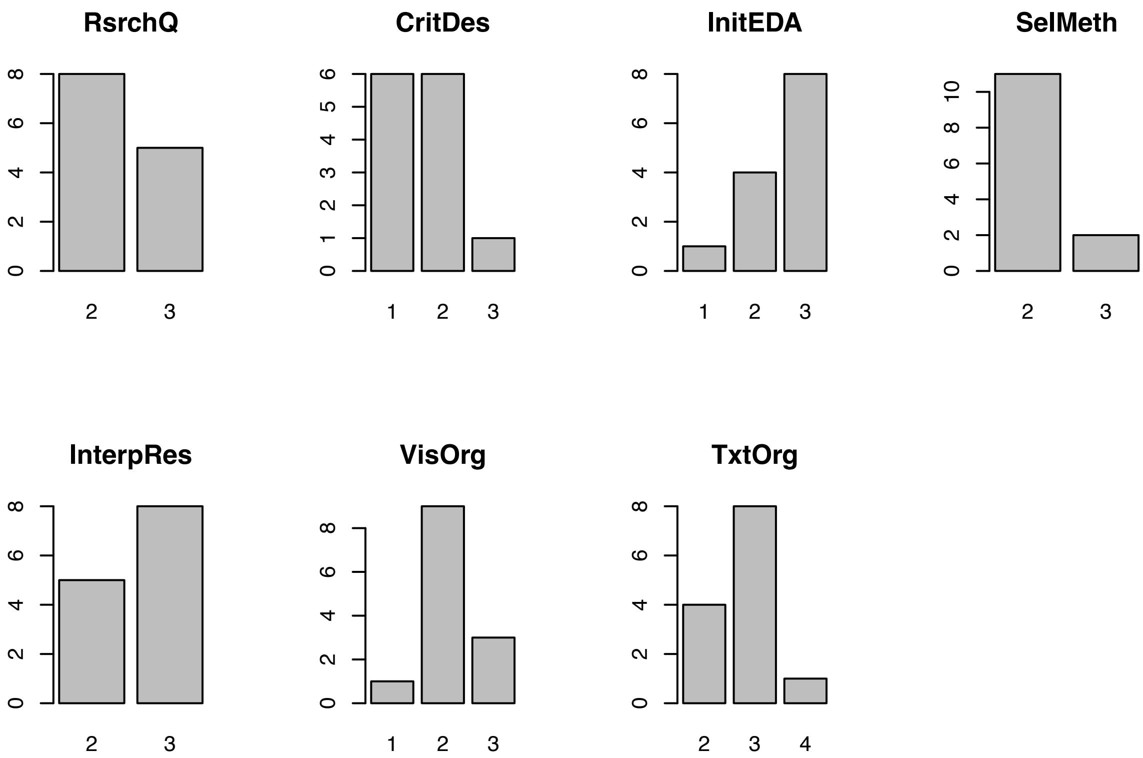

```
plot(all_raters_rubrics$RsrchQ,
     main="RsrchQ")
plot(all_raters_rubrics$CritDes,
     main="CritDes")
plot(all_raters_rubrics$InitEDA,
     main="InitEDA")
plot(all_raters_rubrics$SelMeth,
     main="SelMeth")
plot(all_raters_rubrics$InterpRes,
     main="InterpRes")
plot(all_raters_rubrics$VisOrg,
     main="VisOrg")
plot(all_raters_rubrics$TxtOrg,
     main="TxtOrg")


#Artifacts seen by all raters grouped by rubric and rater
all_raters_1 = ratings %>%
  filter(Repeated==1, Rater==1)
all_raters_1_rubrics = all_raters_1[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```



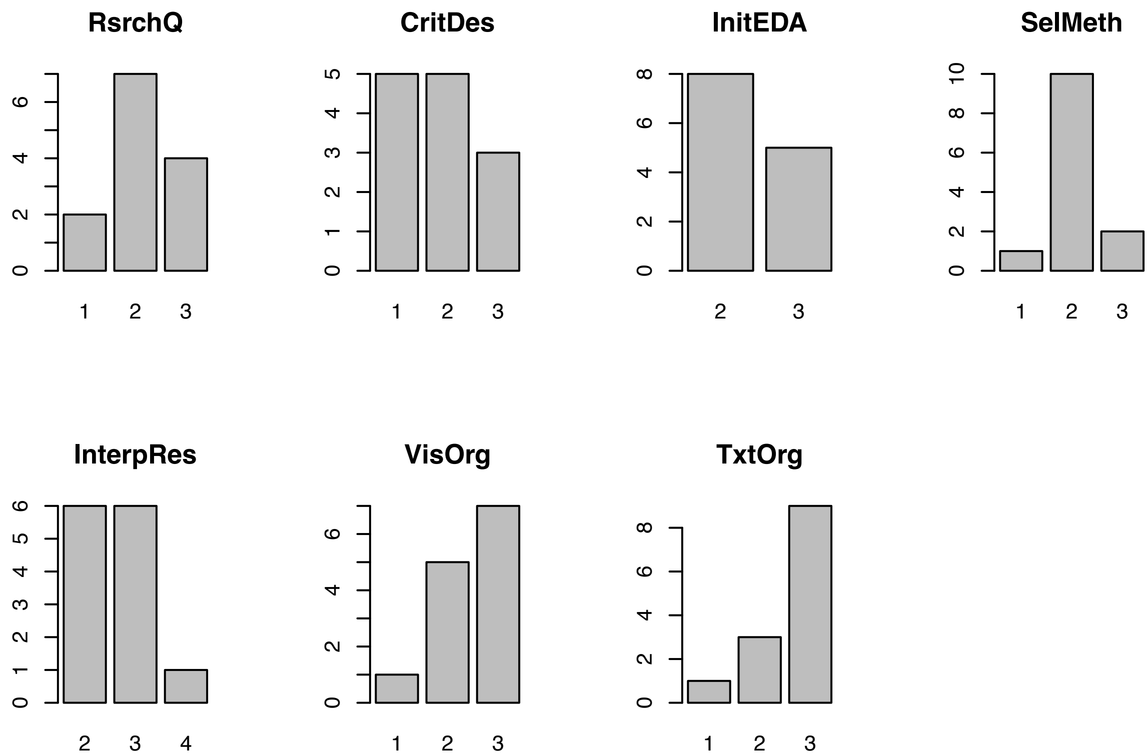

```
plot(all_raters_1_rubrics$RsrchQ,
     main="RsrchQ")
plot(all_raters_1_rubrics$CritDes,
     main="CritDes")
plot(all_raters_1_rubrics$InitEDA,
     main="InitEDA")
plot(all_raters_1_rubrics$SelMeth,
     main="SelMeth")
plot(all_raters_1_rubrics$InterpRes,
     main="InterpRes")
plot(all_raters_1_rubrics$VisOrg,
     main="VisOrg")
plot(all_raters_1_rubrics$TxtOrg,
     main="TxtOrg")


all_raters_2 = ratings %>%
  filter(Repeated==1, Rater==2)
all_raters_2_rubrics = all_raters_2[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```

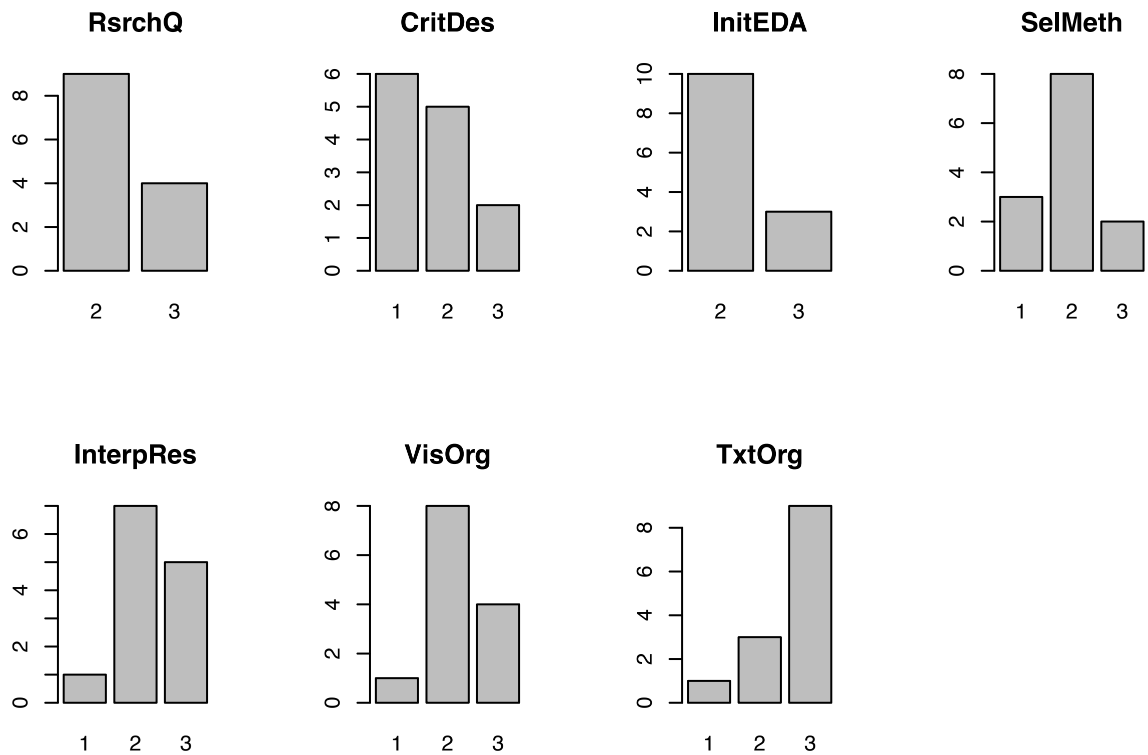



```
plot(all_raters_2_rubrics$RsrchQ,
     main="RsrchQ")
plot(all_raters_2_rubrics$CritDes,
     main="CritDes")
plot(all_raters_2_rubrics$InitEDA,
     main="InitEDA")
plot(all_raters_2_rubrics$SelMeth,
     main="SelMeth")
plot(all_raters_2_rubrics$InterpRes,
     main="InterpRes")
plot(all_raters_2_rubrics$VisOrg,
     main="VisOrg")
plot(all_raters_2_rubrics$TxtOrg,
     main="TxtOrg")


all_raters_3 = ratings %>%
  filter(Repeated==1, Rater==3)
all_raters_3_rubrics = all_raters_3[ , c(7:13)] %>%
  mutate(RsrchQ = as.factor(RsrchQ),
         CritDes = as.factor(CritDes),
         InitEDA = as.factor(InitEDA),
         SelMeth = as.factor(SelMeth),
         InterpRes = as.factor(InterpRes),
         VisOrg = as.factor(VisOrg),
         TxtOrg = as.factor(TxtOrg))
par(mfrow=c(2,4))
```




```
plot(all_raters_3_rubrics$RsrchQ,
     main="RsrchQ")
plot(all_raters_3_rubrics$CritDes,
     main="CritDes")
plot(all_raters_3_rubrics$InitEDA,
     main="InitEDA")
plot(all_raters_3_rubrics$SelMeth,
     main="SelMeth")
plot(all_raters_3_rubrics$InterpRes,
     main="InterpRes")
plot(all_raters_3_rubrics$VisOrg,
     main="VisOrg")
plot(all_raters_3_rubrics$TxtOrg,
     main="TxtOrg")
```



######################## SECTION 2 ########################

```r
#Artifacts Seen by All Raters
all_raters_tall = tall_ratings %>%
  filter(Repeated==1) #%>%
  #mutate(Rater = as.factor(Rater))

rq_ratings = all_raters_tall %>%
  filter(Rubric == "RsrchQ")
ar_rq = lmer(Rating ~ 1+
               (1|Artifact),
             data = rq_ratings)
summary(ar_rq)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: rq_ratings
##
## REML criterion at convergence: 66.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3025 -0.5987 -0.3276  0.9696  1.6472
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05983  0.2446
##  Residual             0.25641  0.5064
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
```

```
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1057   21.59
```

```r
rq_icc = 0.05983/(0.05983+0.25641)

cd_ratings = all_raters_tall %>%
  filter(Rubric == "CritDes")
ar_cd = lmer(Rating ~ 1+
               (1|Artifact),
             data = cd_ratings)
summary(ar_cd)
```

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

```r
cd_icc = 0.3091/(0.3091+0.2308)

ieda_ratings = all_raters_tall %>%
  filter(Rubric == "InitEDA")
ar_ieda = lmer(Rating ~ 1+
               (1|Artifact),
             data = ieda_ratings)
summary(ar_ieda)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: ieda_ratings
##
## REML criterion at convergence: 56.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1670 -0.2504 -0.2504  0.4006  1.6663
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1496   0.3867
##  Residual             0.1538   0.3922
```

```
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.3846     0.1243   19.18
```

```r
ieda_icc = 0.1496/(0.1496+0.1538)

sm_ratings = all_raters_tall %>%
  filter(Rubric == "SelMeth")
ar_sm = lmer(Rating ~ 1+
               (1|Artifact),
             data = sm_ratings)
summary(ar_sm)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: sm_ratings
##
## REML criterion at convergence: 50.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.11366 -0.03357 -0.03357  0.62101  2.04652
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.1396   0.3736
##  Residual             0.1282   0.3581
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.0513     0.1184   17.32
```

```r
sm_icc = 0.1396/(0.1396+0.1282)

ir_ratings = all_raters_tall %>%
  filter(Rubric == "InterpRes")
ar_ir = lmer(Rating ~ 1+
               (1|Artifact),
             data = ir_ratings)
summary(ar_ir)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: ir_ratings
##
## REML criterion at convergence: 71.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0965 -0.8061  0.4844  0.7806  2.6635
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08405  0.2899
##  Residual             0.28205  0.5311
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    2.513      0.117   21.47
```

```
ir_icc = 0.08405/(0.08405+0.28205)

vo_ratings = all_raters_tall %>%
  filter(Rubric == "VisOrg")
ar_vo = lmer(Rating ~ 1+
               (1|Artifact),
             data = vo_ratings)
summary(ar_vo)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: vo_ratings
##
## REML criterion at convergence: 60.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5168 -0.7176 -0.1341  0.3414  1.7241
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.2236   0.4729
##  Residual             0.1538   0.3922
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.2821     0.1454   15.69
```

```
vo_icc = 0.2236/(0.2236+0.1538)

to_ratings = all_raters_tall %>%
  filter(Rubric == "TxtOrg")
ar_to = lmer(Rating ~ 1+
               (1|Artifact),
             data = to_ratings)
summary(ar_to)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: to_ratings
##
## REML criterion at convergence: 74.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.6943 -0.7698  0.3849  0.3849  2.5019
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.05556  0.2357
##  Residual             0.33333  0.5774
## Number of obs: 39, groups:  Artifact, 13
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6667     0.1132   23.55
```

```
to_icc = 0.05556/(0.05556+0.33333)

ar_icc = rbind(rq_icc,
      cd_icc,
      ieda_icc,
      sm_icc,
      ir_icc,
      vo_icc,
      to_icc)
```

```
#exact agreement rates among raters
#raters 1 and 2
#RsrchQ
all_raters = ratings %>%
  filter(Repeated==1)

raters_1_2_rq = data.frame(r1=all_raters$RsrchQ[all_raters$Rater==1],
                    r2=all_raters$RsrchQ[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_rq)
#with(raters_1_2_rq, table(r1,r2))
r1 = factor(raters_1_2_rq$r1, levels=1:4)
r2 = factor(raters_1_2_rq$r2, levels=1:4)
t12_rq = table(r1, r2)/13

#CritDes
raters_1_2_cd = data.frame(r1=all_raters$CritDes[all_raters$Rater==1],
                    r2=all_raters$CritDes[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_cd)
#with(raters_1_2_cd, table(r1,r2))
r1 = factor(raters_1_2_cd$r1, levels=1:4)
r2 = factor(raters_1_2_cd$r2, levels=1:4)
t12_cd = table(r1, r2)/13

#InitEDA
raters_1_2_ieda = data.frame(r1=all_raters$InitEDA[all_raters$Rater==1],
                    r2=all_raters$InitEDA[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_ieda)
```

```
#with(raters_1_2_ieda, table(r1,r2))
r1 = factor(raters_1_2_ieda$r1, levels=1:4)
r2 = factor(raters_1_2_ieda$r2, levels=1:4)
t12_ieda = table(r1, r2)/13

#SelMeth
raters_1_2_sm = data.frame(r1=all_raters$SelMeth[all_raters$Rater==1],
                    r2=all_raters$SelMeth[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_sm)
r1 = factor(raters_1_2_sm$r1, levels=1:4)
r2 = factor(raters_1_2_sm$r2, levels=1:4)
t12_sm = table(r1, r2)/13

#InterpRes
raters_1_2_ir = data.frame(r1=all_raters$InterpRes[all_raters$Rater==1],
                    r2=all_raters$InterpRes[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_ir)
r1 = factor(raters_1_2_ir$r1, levels=1:4)
r2 = factor(raters_1_2_ir$r2, levels=1:4)
t12_ir = table(r1, r2)/13

#VisOrg
raters_1_2_vo = data.frame(r1=all_raters$VisOrg[all_raters$Rater==1],
                    r2=all_raters$VisOrg[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_vo)
r1 = factor(raters_1_2_vo$r1, levels=1:4)
r2 = factor(raters_1_2_vo$r2, levels=1:4)
t12_vo = table(r1, r2)/13

#TxtOrg
raters_1_2_to = data.frame(r1=all_raters$TxtOrg[all_raters$Rater==1],
                    r2=all_raters$TxtOrg[all_raters$Rater==2],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a2=all_raters$Artifact[all_raters$Rater==2])
#View(raters_1_2_to)
r1 = factor(raters_1_2_to$r1, levels=1:4)
r2 = factor(raters_1_2_to$r2, levels=1:4)
t12_to = table(r1, r2)/13
```

```
#exact agreement rates among raters
#raters 1 and 3
#RsrchQ
all_raters = ratings %>%
  filter(Repeated==1)

raters_1_3_rq = data.frame(r1=all_raters$RsrchQ[all_raters$Rater==1],
                    r3=all_raters$RsrchQ[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
```

```
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_rq)
#with(raters_1_3_rq, table(r1,r3))
r1 = factor(raters_1_3_rq$r1, levels=1:4)
r3 = factor(raters_1_3_rq$r3, levels=1:4)
t13_rq = table(r1, r3)/13

#CritDes
raters_1_3_cd = data.frame(r1=all_raters$CritDes[all_raters$Rater==1],
                    r3=all_raters$CritDes[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_cd)
r1 = factor(raters_1_3_cd$r1, levels=1:4)
r3 = factor(raters_1_3_cd$r3, levels=1:4)
t13_cd = table(r1, r3)/13

#InitEDA
raters_1_3_ieda = data.frame(r1=all_raters$InitEDA[all_raters$Rater==1],
                    r3=all_raters$InitEDA[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_ieda)
r1 = factor(raters_1_3_ieda$r1, levels=1:4)
r3 = factor(raters_1_3_ieda$r3, levels=1:4)
t13_ieda = table(r1, r3)/13

#SelMeth
raters_1_3_sm = data.frame(r1=all_raters$SelMeth[all_raters$Rater==1],
                    r3=all_raters$SelMeth[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_sm)
r1 = factor(raters_1_3_sm$r1, levels=1:4)
r3 = factor(raters_1_3_sm$r3, levels=1:4)
t13_sm = table(r1, r3)/13

#InterpRes
raters_1_3_ir = data.frame(r1=all_raters$InterpRes[all_raters$Rater==1],
                    r3=all_raters$InterpRes[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_ir)
r1 = factor(raters_1_3_ir$r1, levels=1:4)
r3 = factor(raters_1_3_ir$r3, levels=1:4)
t13_ir = table(r1, r3)/13

#VisOrg
raters_1_3_vo = data.frame(r1=all_raters$VisOrg[all_raters$Rater==1],
                    r3=all_raters$VisOrg[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_vo)
```

```r
r1 = factor(raters_1_3_vo$r1, levels=1:4)
r3 = factor(raters_1_3_vo$r3, levels=1:4)
t13_vo = table(r1, r3)/13

#TxtOrg
raters_1_3_to = data.frame(r1=all_raters$TxtOrg[all_raters$Rater==1],
                    r3=all_raters$TxtOrg[all_raters$Rater==3],
                    a1=all_raters$Artifact[all_raters$Rater==1],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_1_3_to)
r1 = factor(raters_1_3_to$r1, levels=1:4)
r3 = factor(raters_1_3_to$r3, levels=1:4)
t13_to = table(r1, r3)/13
```

```r
#exact agreement rates among raters
#raters 2 and 3
#RsrchQ
all_raters = ratings %>%
  filter(Repeated==1)

raters_2_3_rq = data.frame(r2=all_raters$RsrchQ[all_raters$Rater==2],
                    r3=all_raters$RsrchQ[all_raters$Rater==3],
                    a2=all_raters$Artifact[all_raters$Rater==2],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_rq)
r2 = factor(raters_2_3_rq$r2, levels=1:4)
r3 = factor(raters_2_3_rq$r3, levels=1:4)
t23_rq = table(r2, r3)/13

#CritDes
raters_2_3_cd = data.frame(r2=all_raters$CritDes[all_raters$Rater==2],
                    r3=all_raters$CritDes[all_raters$Rater==3],
                    a2=all_raters$Artifact[all_raters$Rater==2],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_cd)
r2 = factor(raters_2_3_cd$r2, levels=1:4)
r3 = factor(raters_2_3_cd$r3, levels=1:4)
t23_cd = table(r2, r3)/13

#InitEDA
raters_2_3_ieda = data.frame(r2=all_raters$InitEDA[all_raters$Rater==2],
                    r3=all_raters$InitEDA[all_raters$Rater==3],
                    a2=all_raters$Artifact[all_raters$Rater==2],
                    a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_ieda)
r2 = factor(raters_2_3_ieda$r2, levels=1:4)
r3 = factor(raters_2_3_ieda$r3, levels=1:4)
t23_ieda = table(r2, r3)/13

#SelMeth
raters_2_3_sm = data.frame(r2=all_raters$SelMeth[all_raters$Rater==2],
                    r3=all_raters$SelMeth[all_raters$Rater==3],
                    a2=all_raters$Artifact[all_raters$Rater==2],
                    a3=all_raters$Artifact[all_raters$Rater==3])
```

```r
#View(raters_2_3_sm)
r2 = factor(raters_2_3_sm$r2, levels=1:4)
r3 = factor(raters_2_3_sm$r3, levels=1:4)
t23_sm = table(r2, r3)/13

#InterpRes
raters_2_3_ir = data.frame(r2=all_raters$InterpRes[all_raters$Rater==2],
                     r3=all_raters$InterpRes[all_raters$Rater==3],
                     a2=all_raters$Artifact[all_raters$Rater==2],
                     a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_ir)
r2 = factor(raters_2_3_ir$r2, levels=1:4)
r3 = factor(raters_2_3_ir$r3, levels=1:4)
t23_ir = table(r2, r3)/13

#VisOrg
raters_2_3_vo = data.frame(r2=all_raters$VisOrg[all_raters$Rater==2],
                     r3=all_raters$VisOrg[all_raters$Rater==3],
                     a2=all_raters$Artifact[all_raters$Rater==2],
                     a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_vo)
r2 = factor(raters_2_3_vo$r2, levels=1:4)
r3 = factor(raters_2_3_vo$r3, levels=1:4)
t23_vo = table(r2, r3)/13

#TxtOrg
raters_2_3_to = data.frame(r2=all_raters$TxtOrg[all_raters$Rater==2],
                     r3=all_raters$TxtOrg[all_raters$Rater==3],
                     a2=all_raters$Artifact[all_raters$Rater==2],
                     a3=all_raters$Artifact[all_raters$Rater==3])
#View(raters_2_3_to)
r2 = factor(raters_2_3_to$r2, levels=1:4)
r3 = factor(raters_2_3_to$r3, levels=1:4)
t23_to = table(r2, r3)/13
```

```r
#All Artifacts

rq_ratings = tall_ratings %>%
  filter(Rubric == "RsrchQ")
ar_rq = lmer(Rating ~ 1+
               (1|Artifact),
             data = rq_ratings)
summary(ar_rq)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: rq_ratings
##
## REML criterion at convergence: 211.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2748 -0.5365 -0.3780  0.9626  2.4617
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.07372  0.2715
##  Residual             0.27797  0.5272
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.35790    0.05774   40.84
```

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

cd_ratings = tall_ratings %>%
  filter(Rubric == "CritDes")
ar_cd = lmer(Rating ~ 1+
               (1|Artifact),
             data = cd_ratings)
summary(ar_cd)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: cd_ratings
##
## REML criterion at convergence: 279.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.01120 -0.61076  0.06182  0.73440  2.06404
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.4888   0.6992
##  Residual             0.2409   0.4908
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.90809    0.08779   21.73
```

```
cd_icc = 0.4888/(0.4888+0.2409)

ieda_ratings = tall_ratings %>%
  filter(Rubric == "InitEDA")
ar_ieda = lmer(Rating ~ 1+
               (1|Artifact),
             data = ieda_ratings)
summary(ar_ieda)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: ieda_ratings
##
## REML criterion at convergence: 240.8
##
## Scaled residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.8923 -0.3451 -0.1454  0.4250  1.6015
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3628   0.6023
##  Residual             0.1655   0.4068
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44815    0.07479   32.73
```

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

sm_ratings = tall_ratings %>%
  filter(Rubric == "SelMeth")
ar_sm = lmer(Rating ~ 1+
               (1|Artifact),
             data = sm_ratings)
summary(ar_sm)
```

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

```
sm_icc = 0.1108/(0.1108+0.1240)

ir_ratings = tall_ratings %>%
  filter(Rubric == "InterpRes")
ar_ir = lmer(Rating ~ 1+
               (1|Artifact),
             data = ir_ratings)
summary(ar_ir)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: ir_ratings
##
```

```
## REML criterion at convergence: 217.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1448 -0.6998  0.5175  0.7452  2.6532
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.08219  0.2867
##  Residual             0.29136  0.5398
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.48427    0.05962   41.67
```

```r
ir_icc = 0.08219/(0.08219+0.29136)

vo_ratings = tall_ratings %>%
  filter(Rubric == "VisOrg")
ar_vo = lmer(Rating ~ 1+
               (1|Artifact),
             data = vo_ratings)
summary(ar_vo)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: vo_ratings
##
## REML criterion at convergence: 227.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5894 -0.3772 -0.1628  0.4796  1.6336
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.3063   0.5535
##  Residual             0.1588   0.3985
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.44023    0.07003   34.84
```

```r
vo_icc = 0.3063/(0.3063+0.1588)

to_ratings = tall_ratings %>%
  filter(Rubric == "TxtOrg")
ar_to = lmer(Rating ~ 1+
               (1|Artifact),
             data = to_ratings)
summary(ar_to)
```

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

```
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: to_ratings
##
## REML criterion at convergence: 249
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3638 -0.7641  0.3836  0.5278  2.4094
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Artifact (Intercept) 0.09145  0.3024
##  Residual             0.39503  0.6285
## Number of obs: 117, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.59144    0.06764   38.31
```

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

aa_icc = rbind(rq_icc,
      cd_icc,
      ieda_icc,
      sm_icc,
      ir_icc,
      vo_icc,
      to_icc)
```

```
#comparing ICC for all artifacts and subset of artifacts seen by all raters
icc = data.frame(cbind(ar_icc, aa_icc)) %>%
  mutate(all_raters= X1,
         X1 = NULL,
         all_artifacts = X2,
         X2 = NULL)

kable(icc, caption="Intraclass Correlation")
```

Table 1: Intraclass Correlation

	all_raters	all_artifacts
rq_icc	0.1891918	0.2096164
cd_icc	0.5725134	0.6698643
ieda_icc	0.4930784	0.6867310
sm_icc	0.5212845	0.4718910
ir_icc	0.2295821	0.2200241
vo_icc	0.5924748	0.6585680
to_icc	0.1428682	0.1879831

```
#Percent Exact Agreement Between Raters
exact_agreement = as.data.frame(cbind(rbind(sum(diag(t12_rq)),
sum(diag(t12_cd)),
sum(diag(t12_ieda)),
sum(diag(t12_sm)),
```

```
sum(diag(t12_ir)),
sum(diag(t12_vo)),
sum(diag(t12_to))),
rbind(sum(diag(t13_rq)),
sum(diag(t13_cd)),
sum(diag(t13_ieda)),
sum(diag(t13_sm)),
sum(diag(t13_ir)),
sum(diag(t13_vo)),
sum(diag(t13_to))),
rbind(sum(diag(t23_rq)),
sum(diag(t23_cd)),
sum(diag(t23_ieda)),
sum(diag(t23_sm)),
sum(diag(t23_ir)),
sum(diag(t23_vo)),
sum(diag(t23_to))))) %>%
  mutate(rubric = c("rq", "cd", "ieda", "sm", "ir", "vo", "to"),
         raters_12 = V1,
         V1 = NULL,
         raters_13 = V2,
         V2 = NULL,
         raters_23 = V3,
         V3 = NULL)

kable(exact_agreement, caption = "Percent Exact Agreement")
```

Table 2: Percent Exact Agreement

rubric	raters_12	raters_13	raters_23
rq	0.3846154	0.7692308	0.5384615
cd	0.5384615	0.6153846	0.6923077
ieda	0.6923077	0.5384615	0.8461538
sm	0.9230769	0.6153846	0.6923077
ir	0.6153846	0.5384615	0.6153846
vo	0.5384615	0.7692308	0.7692308
to	0.6923077	0.6153846	0.5384615

######################## SECTION 3 ########################

```
#selecting fixed effects
simple_mod = lmer(Rating ~ (0 + Rubric | Artifact),
                  data = tall_ratings)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

```
summary(simple_mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact)
```

```
##    Data: tall_ratings
##
## REML criterion at convergence: 1484.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0286 -0.5036 -0.0755  0.5140  3.7802
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.6404   0.8003
##           RubricInitEDA   0.3784   0.6151   0.27
##           RubricInterpRes 0.2526   0.5026   0.02 0.79
##           RubricRsrchQ    0.1738   0.4169   0.40 0.51 0.74
##           RubricSelMeth   0.1033   0.3214   0.58 0.39 0.42 0.29
##           RubricTxtOrg    0.3951   0.6286   0.04 0.69 0.80 0.64 0.25
##           RubricVisOrg    0.3132   0.5597   0.19 0.79 0.77 0.61 0.30 0.80
##  Residual                 0.1941   0.4406
## Number of obs: 819, groups:  Artifact, 91
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.24615    0.04045   55.53
## optimizer (nloptwrap) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

```
medium_mod = lmer(Rating ~ Rater +
                  Rubric +
                  (0 + Rubric | Artifact),
                   data = tall_ratings)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0046949 (tol = 0.002, component 1)
```

```
kable(summary(medium_mod)$coefficients)
```

	Estimate	Std. Error	t value
(Intercept)	1.9673124	0.0945268	20.8122185
Rater2	0.0008391	0.0553345	0.0151639
Rater3	-0.1662660	0.0553345	-3.0047437
RubricInitEDA	0.5385874	0.0943645	5.7075244
RubricInterpRes	0.5768621	0.0994519	5.8004150
RubricRsrchQ	0.4506140	0.0861883	5.2282490
RubricSelMeth	0.1550413	0.0915983	1.6926224
RubricTxtOrg	0.6833084	0.0978861	6.9806442
RubricVisOrg	0.5170880	0.0976362	5.2960714

```
full_mod = lmer(Rating ~ Rater +
                  Semester +
                  Sex +
                  Repeated +
                  Rubric +
                  (0 + Rubric | Artifact),
```

```
                   data = tall_ratings)
```

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

```
summary(full_mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + Semester + Sex + Repeated + Rubric + (0 + Rubric |
##     Artifact)
##    Data: tall_ratings
##
## REML criterion at convergence: 1438.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1173 -0.5074 -0.0266  0.5215  3.7742
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.54119  0.7357
##           RubricInitEDA   0.34775  0.5897   0.47
##           RubricInterpRes 0.17310  0.4161   0.23 0.76
##           RubricRsrchQ    0.16758  0.4094   0.59 0.45 0.72
##           RubricSelMeth   0.06744  0.2597   0.40 0.61 0.75 0.42
##           RubricTxtOrg    0.25874  0.5087   0.35 0.62 0.74 0.56 0.67
##           RubricVisOrg    0.25333  0.5033   0.34 0.75 0.68 0.53 0.41 0.78
##  Residual                 0.18988  0.4358
## Number of obs: 819, groups:  Artifact, 91
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      2.823556   0.388203   7.273
## Rater2           0.002947   0.054950   0.054
## Rater3          -0.174527   0.055110  -3.167
## SemesterS19     -0.174664   0.087784  -1.990
## SexF            -0.803550   0.383604  -2.095
## SexM            -0.793346   0.382616  -2.073
## Repeated        -0.074274   0.098449  -0.754
## RubricInitEDA    0.539223   0.094364   5.714
## RubricInterpRes  0.576874   0.099409   5.803
## RubricRsrchQ     0.454182   0.086215   5.268
## RubricSelMeth    0.160365   0.092622   1.731
## RubricTxtOrg     0.683479   0.097598   7.003
## RubricVisOrg     0.518469   0.097702   5.307
```

```
##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
medium_int_mod = lmer(Rating ~ Rater +
                  Rubric +
                    Rater:Semester +
```

```r
                   Rater:Sex +
                 (0 + Rubric | Artifact),
                 data = tall_ratings)
```

```
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```r
summary(medium_int_mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ Rater + Rubric + Rater:Semester + Rater:Sex + (0 + Rubric |
##     Artifact)
##    Data: tall_ratings
##
## REML criterion at convergence: 1441.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2065 -0.4908 -0.0396  0.5045  3.7490
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.55633  0.7459
##           RubricInitEDA   0.35796  0.5983   0.48
##           RubricInterpRes 0.17457  0.4178   0.24 0.76
##           RubricRsrchQ    0.17090  0.4134   0.59 0.45 0.71
##           RubricSelMeth   0.07067  0.2658   0.43 0.62 0.75 0.42
##           RubricTxtOrg    0.25999  0.5099   0.35 0.62 0.69 0.56 0.66
##           RubricVisOrg    0.26915  0.5188   0.36 0.74 0.69 0.54 0.44 0.76
##  Residual                 0.18620  0.4315
## Number of obs: 819, groups:  Artifact, 91
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)         2.046963   0.111940  18.286
## Rater2              0.001298   0.083923   0.015
## Rater3              0.600406   0.387151   1.551
## RubricInitEDA       0.538546   0.094351   5.708
## RubricInterpRes     0.576842   0.099455   5.800
## RubricRsrchQ        0.450610   0.086066   5.236
## RubricSelMeth       0.153926   0.092035   1.672
## RubricTxtOrg        0.682446   0.097955   6.967
## RubricVisOrg        0.517458   0.097668   5.298
## Rater1:SemesterS19 -0.092714   0.114925  -0.807
## Rater2:SemesterS19 -0.147104   0.115060  -1.278
## Rater3:SemesterS19 -0.258890   0.115668  -2.238
## Rater1:SexF        -0.091931   0.105682  -0.870
## Rater2:SexF        -0.072566   0.103499  -0.701
## Rater3:SexF        -0.728895   0.388028  -1.878
## Rater3:SexM        -0.848566   0.386802  -2.194

##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```
full_int_mod = lmer(Rating ~ Rater +
                    Semester +
                    Sex +
                    Repeated +
                    Rubric +
                    Rater:Semester +
                      Rater:Sex +
                    (0 + Rubric | Artifact),
                     data = tall_ratings)
```

```
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```
summary(full_int_mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## Rating ~ Rater + Semester + Sex + Repeated + Rubric + Rater:Semester +
##     Rater:Sex + (0 + Rubric | Artifact)
##    Data: tall_ratings
##
## REML criterion at convergence: 1443.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1943 -0.4931 -0.0424  0.4990  3.7645
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.55336  0.7439
##           RubricInitEDA   0.35859  0.5988   0.47
##           RubricInterpRes 0.17893  0.4230   0.24 0.76
##           RubricRsrchQ    0.17116  0.4137   0.59 0.45 0.71
##           RubricSelMeth   0.07293  0.2700   0.41 0.61 0.75 0.42
##           RubricTxtOrg    0.26573  0.5155   0.35 0.62 0.70 0.57 0.67
##           RubricVisOrg    0.26580  0.5156   0.35 0.74 0.69 0.53 0.43 0.77
##  Residual                 0.18599  0.4313
## Number of obs: 819, groups:  Artifact, 91
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)          2.896024   0.397764   7.281
## Rater2               0.001092   0.083979   0.013
## Rater3              -0.247863   0.083988  -2.951
## SemesterS19         -0.099738   0.115451  -0.864
## SexF                -0.929481   0.397795  -2.337
## SexM                -0.837490   0.388851  -2.154
## Repeated            -0.074050   0.100084  -0.740
## RubricInitEDA        0.538198   0.094345   5.705
## RubricInterpRes      0.577455   0.099483   5.805
## RubricRsrchQ         0.453070   0.086167   5.258
## RubricSelMeth        0.159432   0.092507   1.723
## RubricTxtOrg         0.682438   0.097938   6.968
```

```
## RubricVisOrg        0.518686   0.097700   5.309
## Rater2:SemesterS19 -0.051948   0.127968  -0.406
## Rater3:SemesterS19 -0.163952   0.128146  -1.279
## Rater2:SexF         0.022063   0.112353   0.196
## Rater3:SexF         0.213489   0.112678   1.895
```

```
##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients
```

```
fe_anova = anova(simple_mod, medium_mod, full_mod, medium_int_mod, full_int_mod)
```

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

```
anova(medium_mod, medium_int_mod, full_int_mod)
```

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

```
## Data: tall_ratings
## Models:
## medium_mod: Rating ~ Rater + Rubric + (0 + Rubric | Artifact)
## medium_int_mod: Rating ~ Rater + Rubric + Rater:Semester + Rater:Sex + (0 + Rubric | Artifact)
## full_int_mod: Rating ~ Rater + Semester + Sex + Repeated + Rubric + Rater:Semester + Rater:Sex + (0 +
##                 npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## medium_mod        38 1479.8 1658.7 -701.88   1403.8
## medium_int_mod    45 1479.6 1691.5 -694.81   1389.6 14.1395  7    0.04876 *
## full_int_mod      46 1481.1 1697.7 -694.54   1389.1  0.5399  1    0.46248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
BIC(simple_mod)
```

```
## [1] 1685.362
```

```
BIC(medium_mod)
```

```
## [1] 1693.65
```

```
BIC(full_mod)
```

```
## [1] 1720.651
```

```
BIC(medium_int_mod)
```

```
## [1] 1743.382
```

```
BIC(full_int_mod)
```

```
## [1] 1752.36
```

```
#medium_mod produces best results --> best BIC, purpose isn't to predict what rating
#an artifact will receive, want interpretable results (lec. 23, slide 11)
```

```
simpl_mod1 = lmer(Rating ~ Semester + (0 + Rubric | Artifact),
                  data = tall_ratings)
anova(simpl_mod1, simple_mod)$BIC #not significantly different BICs
```

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

```
## [1] 1680.759 1683.437
```

```
med_mod1 = lmer(Rating ~ Rater +
                  Rubric +
                  Repeated +
                  (0 + Rubric | Artifact),
                   data = tall_ratings)

med_mod2 = lmer(Rating ~ Rater +
                  Rubric +
                  Semester +
                  (0 + Rubric | Artifact),
                   data = tall_ratings)

med_mod3 = lmer(Rating ~ Rater +
                  Rubric +
                  Sex +
                  (0 + Rubric | Artifact),
                   data = tall_ratings)
```

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

```
anova(medium_mod, med_mod1, med_mod2, med_mod3)$BIC
```

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

## [1] 1658.664 1664.959 1661.084 1666.580
```

```
BIC(medium_mod)
```

```
## [1] 1693.65
```

```
BIC(med_mod1)
```

```
## [1] 1702.681
```

```
BIC(med_mod2)
```

```
## [1] 1699.252
```

```
BIC(med_mod3)
```

```
## [1] 1705.022
```

```
#selecting random effects
simple_re_mod = lmer(Rating ~ Rater +
                  Rubric +
                  (0 + Rubric | Artifact),
                   data = tall_ratings) #medium_mod
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0046949 (tol = 0.002, component 1)
```

```
full_re_mod = lmer(Rating ~ Rater +
                     Rubric +
                    (0 + Rubric | Artifact) +
                    (0 + Rater | Artifact),
                   data = tall_ratings)
```

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

```r
re_anova = anova(simple_re_mod, full_re_mod) #full_re_mod preferred
```

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

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

```r
kable(summary(full_re_mod)$coefficients)
```

	Estimate	Std. Error	t value
(Intercept)	1.9629783	0.0927901	21.1550312
Rater2	0.0069414	0.0783864	0.0885534
Rater3	-0.1608235	0.0676811	-2.3761947
RubricInitEDA	0.5353223	0.0941189	5.6877217
RubricInterpRes	0.5746624	0.0993261	5.7856135
RubricRsrchQ	0.4488698	0.0853845	5.2570407
RubricSelMeth	0.1508231	0.0915866	1.6467810
RubricTxtOrg	0.6748400	0.0985931	6.8446943
RubricVisOrg	0.5210053	0.0978035	5.3270617

```r
#exact test of random effect (lec. 23, slide 20)
rater_re_mod = lmer(Rating ~ Rater +
                      Rubric +
                     (0 + Rater | Artifact),
                   data = tall_ratings)
```

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

```r
#H_0: simple_re_mod
#H_a: full_re_mod
#model with only random slopes for rater (lec. 23, slide 20): rater_re_mod
#exactRLRT(rater_re_mod, full_re_mod, simple_re_mod)
```

#report and explain Betas –> mean rating in each rubric #report tau^2 from etas –> how much variation on each rubric for each artifact #for etas find high and low etas –> most different from Betas (which is mean) #Calculating Beta values for fixed and random effects of chosen model (full_re_mod)

```r
#fixed effects
beta0 <- fixef(full_re_mod)[1] #Intercept --> Rater1 and CritDes
beta_r2 <- beta0 + fixef(full_re_mod)[2] #Rater2
beta_r3 <- beta0 + fixef(full_re_mod)[3] #Rater3
beta_ieda <- beta0 + fixef(full_re_mod)[4] #InitEDA
beta_ir <- beta0 + fixef(full_re_mod)[5] #InterpRes
beta_rq <- beta0 + fixef(full_re_mod)[6] #RsrchQ
beta_sm <- beta0 + fixef(full_re_mod)[7] #SelMeth
beta_to <- beta0 + fixef(full_re_mod)[8] #TxtOrg
beta_vo <- beta0 + fixef(full_re_mod)[9] #VisOrg

#random effects
eta = ranef(full_re_mod)$Artifact

alpha_cd1 = beta0 + eta[,1] + eta[,8] #CritDes for Rater1
alpha_cd2 = beta0 + beta_r2 + eta[,1] + eta[,9] #CritDes for Rater2
alpha_cd3 = beta0 + beta_r3 + eta[,1] + eta[,10] #CritDes for Rater3
beta_cd1 = beta0
```

```
beta_cd2 = beta0 + beta_r2
beta_cd3 = beta0 + beta_r3

alpha_ieda1 = beta0 + beta_ieda + eta[,2] + eta[,8] #InitEDA for Rater1
alpha_ieda2 = beta0 + beta_r2 + beta_ieda + eta[,2] + eta[,9] #InitEDA for Rater2
alpha_ieda3 = beta0 + beta_r3 + beta_ieda + eta[,2] + eta[,10] #InitEDA for Rater3
beta_ieda1 = beta0 + beta_ieda
beta_ieda2 = beta0 + beta_r2 + beta_ieda
beta_ieda3 = beta0 + beta_r3 + beta_ieda

alpha_ir1 = beta0 + beta_ir + eta[,3] + eta[,8] #InterpRes for Rater1
alpha_ir2 = beta0 + beta_r2 + beta_ir + eta[,3] + eta[,9] #InterpRes for Rater2
alpha_ir3 = beta0 + beta_r3 + beta_ir + eta[,3] + eta[,10] #InterpRes for Rater3
beta_ir1 = beta0 + beta_ir
beta_ir2 = beta0 + beta_r2 + beta_ir
beta_ir3 = beta0 + beta_r3 + beta_ir

alpha_rq1 = beta0 + beta_rq + eta[,4] + eta[,8] #RsrchQ for Rater1
alpha_rq2 = beta0 + beta_r2 + beta_rq + eta[,4] + eta[,9] #RsrchQ for Rater2
alpha_rq3 = beta0 + beta_r3 + beta_rq + eta[,4] + eta[,10] #RsrchQ for Rater3
beta_rq1 = beta0 + beta_rq
beta_rq2 = beta0 + beta_r2 + beta_rq
beta_rq3 = beta0 + beta_r3 + beta_rq

alpha_sm1 = beta0 + beta_sm + eta[,5] + eta[,8] #SelMeth for Rater1
alpha_sm2 = beta0 + beta_r2 + beta_sm + eta[,5] + eta[,9] #SelMeth for Rater2
alpha_sm3 = beta0 + beta_r3 + beta_sm + eta[,5] + eta[,10] #SelMeth for Rater3
beta_sm1 = beta0 + beta_sm
beta_sm2 = beta0 + beta_r2 + beta_sm
beta_sm3 = beta0 + beta_r3 + beta_sm

alpha_to1 = beta0 + beta_to + eta[,6] + eta[,8] #TxtOrg for Rater1
alpha_to2 = beta0 + beta_r2 + beta_to + eta[,6] + eta[,9] #TxtOrg for Rater2
alpha_to3 = beta0 + beta_r3 + beta_to + eta[,6] + eta[,10] #TxtOrg for Rater3
beta_to1 = beta0 + beta_to
beta_to2 = beta0 + beta_r2 + beta_to
beta_to3 = beta0 + beta_r3 + beta_to

alpha_vo1 = beta0 + beta_vo + eta[,7] + eta[,8] #VisOrg for Rater1
alpha_vo2 = beta0 + beta_r2 + beta_vo + eta[,7] + eta[,9] #VisOrg for Rater2
alpha_vo3 = beta0 + beta_r3 + beta_vo + eta[,7] + eta[,10] #VisOrg for Rater3
beta_vo1 = beta0 + beta_vo
beta_vo2 = beta0 + beta_r2 + beta_vo
beta_vo3 = beta0 + beta_r3 + beta_vo

full_re_alphas = as.data.frame(cbind(alpha_cd1,
                       alpha_cd2,
                       alpha_cd3,
                       alpha_ieda1,
                       alpha_ieda2,
                       alpha_ieda3,
                       alpha_ir1,
                       alpha_ir2,
```

```
                       alpha_ir3,
                       alpha_rq1,
                       alpha_rq2,
                       alpha_rq3,
                       alpha_sm1,
                       alpha_sm2,
                       alpha_sm3,
                       alpha_to1,
                       alpha_to2,
                       alpha_to3,
                       alpha_vo1,
                       alpha_vo2,
                       alpha_vo3))
rownames(full_re_alphas) = rownames(eta)

alphas = cbind(rbind(beta_cd1,
                     beta_cd2,
                     beta_cd3,
                     beta_ieda1,
                     beta_ieda2,
                     beta_ieda3,
                     beta_ir1,
                     beta_ir2,
                     beta_ir3,
                     beta_rq1,
                     beta_rq2,
                     beta_rq3,
                     beta_sm1,
                     beta_sm2,
                     beta_sm3,
                     beta_to1,
                     beta_to2,
                     beta_to3,
                     beta_vo1,
                     beta_vo2,
                     beta_vo3),
                     lapply(full_re_alphas, min),
      lapply(full_re_alphas, max))
rownames(alphas) = c("cd1", "cd2", "cd3",
                     "ieda1", "ieda2", "ieda3",
                     "ir1", "ir2", "ir3",
                     "rq1", "rq2", "rq3",
                     "sm1", "sm2", "sm3",
                     "to1", "to2", "to3",
                     "vo1", "vo2", "vo3")
colnames(alphas) = c("Beta", "min(alpha)", "max(alpha)")
kable(alphas)
```

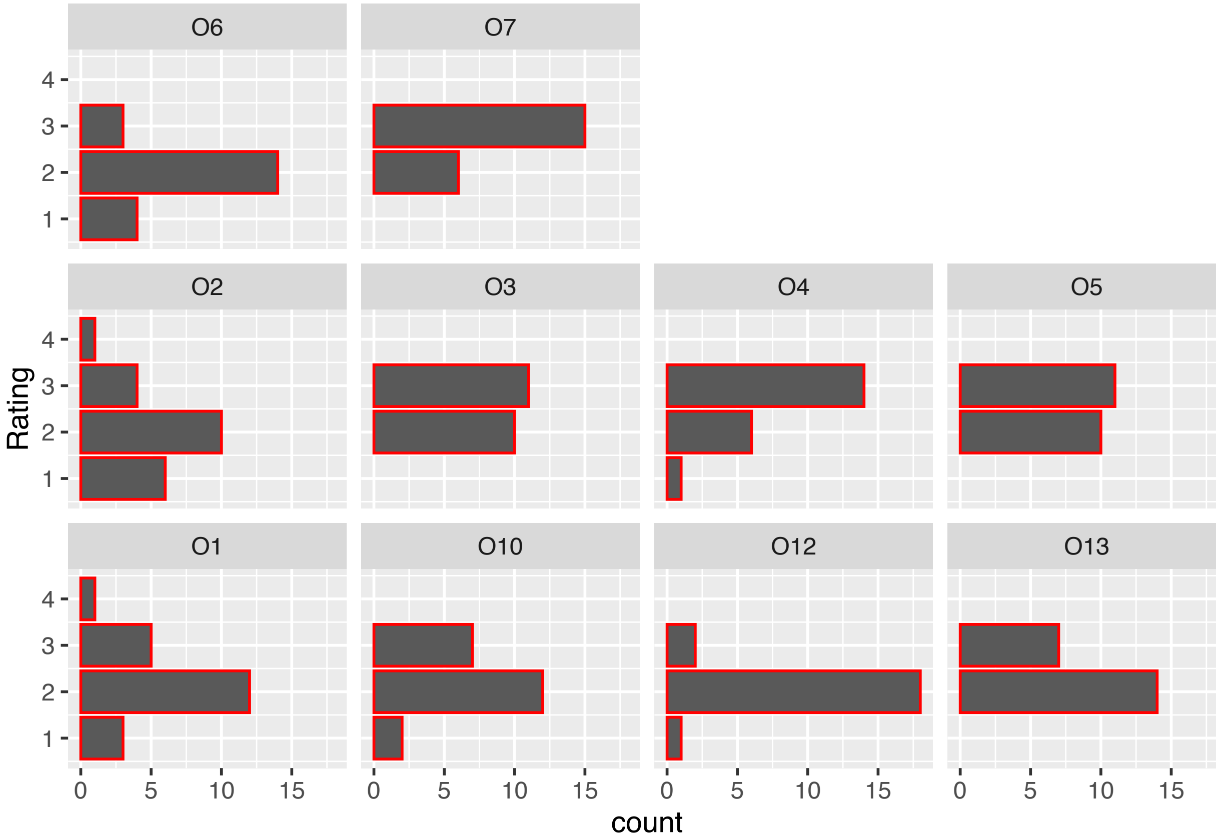
	Beta	min(alpha)	max(alpha)
cd1	1.96297829962909	0.964090709463024	3.35318653481723
cd2	3.93289798735599	2.84191804151764	5.63654001488785
cd3	3.7651331191995	2.75171839371924	5.38339875879339
ieda1	4.46127887235246	3.46372367547626	5.59812337589804

	Beta	min(alpha)	max(alpha)
ieda2	6.43119856007936	5.32232366754462	7.68588959967795
ieda3	6.26343369192287	4.96913361137758	7.58259576705219
ir1	4.50061895114662	3.91854527060323	4.93243797911713
ir2	6.47053863887352	5.52909246246343	7.14759046475785
ir3	6.30277377071703	5.13831704193551	6.94076493093936
rq1	4.37482641217448	3.58052146777698	5.28271271531842
rq2	6.34474609990138	5.44200256407742	7.09587328872512
rq3	6.17698123174489	5.12591726351563	7.09929590648849
sm1	4.07677968780044	3.75115602618428	4.4227085286392
sm2	6.04669937552733	4.9852566706079	6.85672945391781
sm3	5.87893450737085	5.08162855153722	6.50156725247223
to1	4.60079656491439	3.5376863135765	5.48041389551835
to2	6.57071625264128	5.37104511920654	7.59089524794113
to3	6.4029513844848	5.10411123583649	7.52809595642987
vo1	4.44696189039077	3.4246549554829	5.40435860628462
vo2	6.41688157811767	5.18656603471481	7.53780792508445
vo3	6.24911670996118	4.88457920254964	7.28466666898999

######################## SECTION 4 ########################

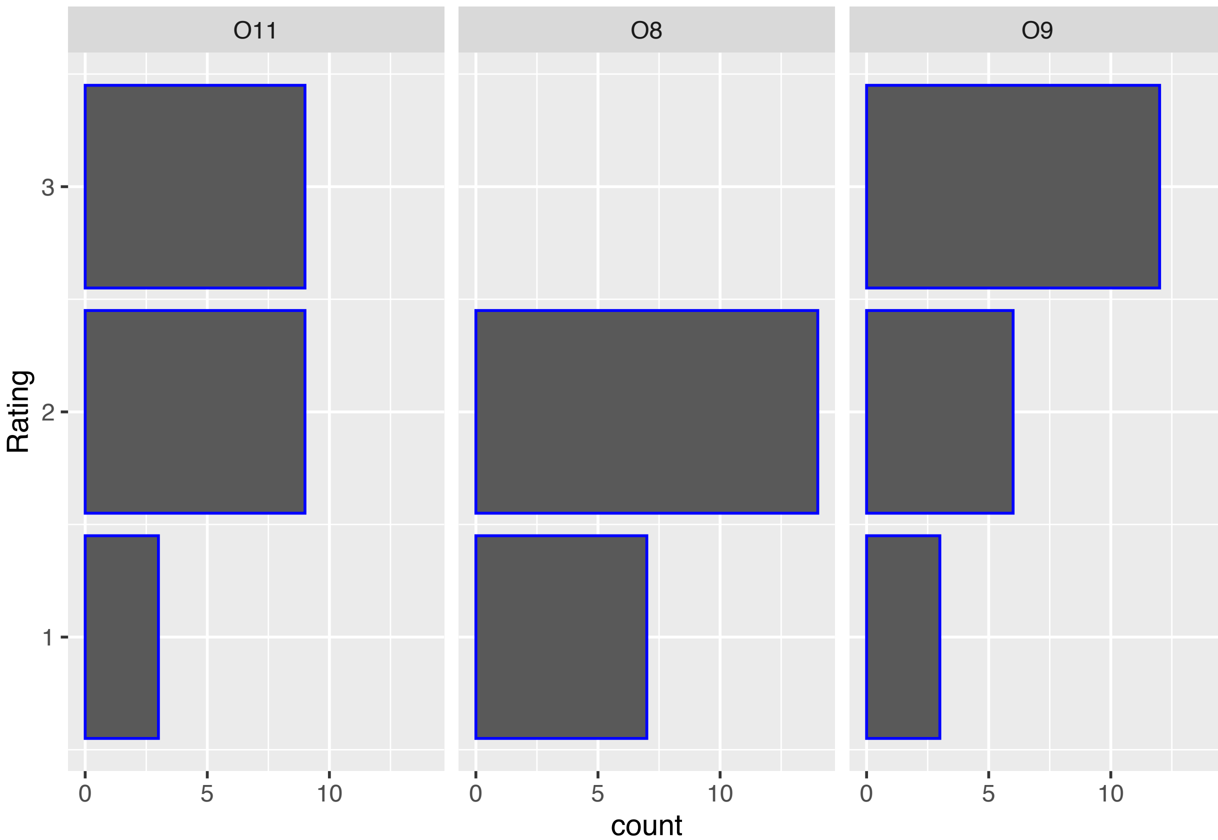
```
#check rubric distributions between semesters
#find rater disagreement
all_raters_fall = tall_ratings %>%
  filter(Repeated == 1, Semester == "F19")
ggplot(all_raters_fall, aes(y=Rating)) +
  facet_wrap( ~ Artifact, as.table=F) +
    geom_bar(pch=1,color="red")
```

```
## Warning: Ignoring unknown parameters: shape
```




```
all_raters_spring = tall_ratings %>%
  filter(Repeated == 1, Semester == "S19")
ggplot(all_raters_spring, aes(y=Rating)) +
  facet_wrap( ~ Artifact, as.table=F) +
    geom_bar(pch=1,color="blue")
```

```
## Warning: Ignoring unknown parameters: shape
```




```
exact_disagreement = as.data.frame(cbind(rbind(
      t12_rq[4,1] + t12_rq[3,1] + t12_rq[4,2] + t12_rq[1,3] + t12_rq[1,4] + t12_rq[2,4],
      t12_cd[4,1] + t12_cd[3,1] + t12_cd[4,2] + t12_cd[1,3] + t12_cd[1,4] + t12_cd[2,4],
      t12_ieda[4,1] + t12_ieda[3,1] + t12_ieda[4,2] + t12_ieda[1,3] + t12_ieda[1,4] + t12_ieda[2,4],
      t12_sm[4,1] + t12_sm[3,1] + t12_sm[4,2] + t12_sm[1,3] + t12_sm[1,4] + t12_sm[2,4],
      t12_ir[4,1] + t12_ir[3,1] + t12_ir[4,2] + t12_ir[1,3] + t12_ir[1,4] + t12_ir[2,4],
      t12_vo[4,1] + t12_vo[3,1] + t12_vo[4,2] + t12_vo[1,3] + t12_vo[1,4] + t12_vo[2,4],
      t12_to[4,1] + t12_to[3,1] + t12_to[4,2] + t12_to[1,3] + t12_to[1,4] + t12_to[2,4]),
      rbind(
      t13_rq[4,1] + t13_rq[3,1] + t13_rq[4,2] + t13_rq[1,3] + t13_rq[1,4] + t13_rq[2,4],
      t13_cd[4,1] + t13_cd[3,1] + t13_cd[4,2] + t13_cd[1,3] + t13_cd[1,4] + t13_cd[2,4],
      t13_ieda[4,1] + t13_ieda[3,1] + t13_ieda[4,2] + t13_ieda[1,3] + t13_ieda[1,4] + t13_ieda[2,4],
      t13_sm[4,1] + t13_sm[3,1] + t13_sm[4,2] + t13_sm[1,3] + t13_sm[1,4] + t13_sm[2,4],
      t13_ir[4,1] + t13_ir[3,1] + t13_ir[4,2] + t13_ir[1,3] + t13_ir[1,4] + t13_ir[2,4],
      t13_vo[4,1] + t13_vo[3,1] + t13_vo[4,2] + t13_vo[1,3] + t13_vo[1,4] + t13_vo[2,4],
      t13_to[4,1] + t13_to[3,1] + t13_to[4,2] + t13_to[1,3] + t13_to[1,4] + t13_to[2,4]),
      rbind(
      t23_rq[4,1] + t23_rq[3,1] + t23_rq[4,2] + t23_rq[1,3] + t23_rq[1,4] + t23_rq[2,4],
      t23_cd[4,1] + t23_cd[3,1] + t23_cd[4,2] + t23_cd[1,3] + t23_cd[1,4] + t23_cd[2,4],
      t23_ieda[4,1] + t23_ieda[3,1] + t23_ieda[4,2] + t23_ieda[1,3] + t23_ieda[1,4] + t23_ieda[2,4],
      t23_sm[4,1] + t23_sm[3,1] + t23_sm[4,2] + t23_sm[1,3] + t23_sm[1,4] + t23_sm[2,4],
      t23_ir[4,1] + t23_ir[3,1] + t23_ir[4,2] + t23_ir[1,3] + t23_ir[1,4] + t23_ir[2,4],
      t23_vo[4,1] + t23_vo[3,1] + t23_vo[4,2] + t23_vo[1,3] + t23_vo[1,4] + t23_vo[2,4],
      t23_to[4,1] + t23_to[3,1] + t23_to[4,2] + t23_to[1,3] + t23_to[1,4] + t23_to[2,4]))) %>%
  mutate(rubric = c("rq", "cd", "ieda", "sm", "ir", "vo", "to"),
         raters_12 = V1,
         V1 = NULL,
```

```r
         raters_13 = V2,
         V2 = NULL,
         raters_23 = V3,
         V3 = NULL)

exact_agreement
```

```
##   rubric raters_12 raters_13 raters_23
## 1     rq 0.3846154 0.7692308 0.5384615
## 2     cd 0.5384615 0.6153846 0.6923077
## 3   ieda 0.6923077 0.5384615 0.8461538
## 4     sm 0.9230769 0.6153846 0.6923077
## 5     ir 0.6153846 0.5384615 0.6153846
## 6     vo 0.5384615 0.7692308 0.7692308
## 7     to 0.6923077 0.6153846 0.5384615
```

```r
kable(exact_disagreement)
```

rubric	raters_12	raters_13	raters_23
rq	0.0769231	0.0000000	0.0000000
cd	0.0769231	0.0000000	0.0000000
ieda	0.0000000	0.0000000	0.0000000
sm	0.0000000	0.0000000	0.0000000
ir	0.0769231	0.0000000	0.0769231
vo	0.0000000	0.0000000	0.0000000
to	0.0769231	0.0769231	0.0000000