Appendices:

Jie Luo

11/30/2020

Contents

Appendix: Initial Data Import & Exploration

Read the data. get a general idea of the variables. Using table and boxplots for each rubric.

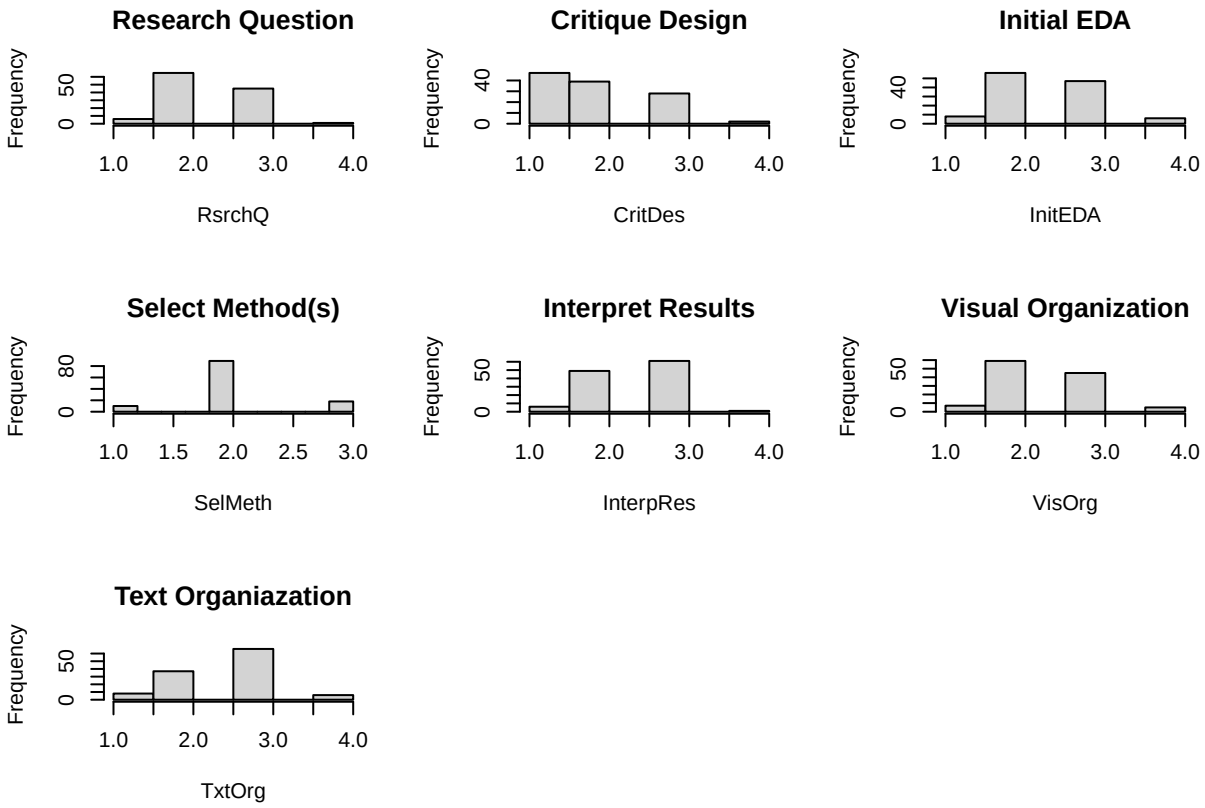
```
library(arm)
library(lme4)
library(plyr)
library(tidyverse)
```

```
library(gtsummary)
rating <- read.csv("ratings.csv")
```

```
subrating <- rating %>%
  drop_na()
```

First, display the histogram from each rubric. Since for each rubric, only integer 1,2,3, and 4 are given. So, the scores are discrete.

```
par(mfrow = c(3,3))
attach(rating)
hist(RsrchQ, main = "Research Question")
hist(CritDes, main = "Critique Design")
hist(InitEDA, main = "Initial EDA")
hist(SelMeth, main = "Select Method(s)")
hist(InterpRes, main = "Interpret Results")
hist(VisOrg, main = "Visual Organization")
hist(TxtOrg, main = "Text Organiazation")
detach()
```



From the plot, we can see that the distributions for each rubric are a little different from each other. Research question, Initial EDA and visual organization have similar distributions. The rest differs from each other. Respectively, for research question, initial EDA, and visual organization, the most frequent score is 2, then to 3. For Critique Design, the most frequent score is 1, then 2 and 3. As for select Method, the most frequent score is 2 and no one has score as 4. The interpret Results and Text organization have the similar pattern that most frequent scores are 3. Then it comes to 2.

To conclude, the Rubric Critique Design tends to have low score(majority is 1). The rest of rubrics tend to have the majority of score be 2 or 3. This might suggest the random effect when considering the model selection

```
library(arsenal)

table_one <- tableby(Rater ~ RsrchQ + CritDes + InitEDA + SelMeth + InterpRes + VisOrg + TxtOrg, data=ra
```

Table.1 Summary statistics

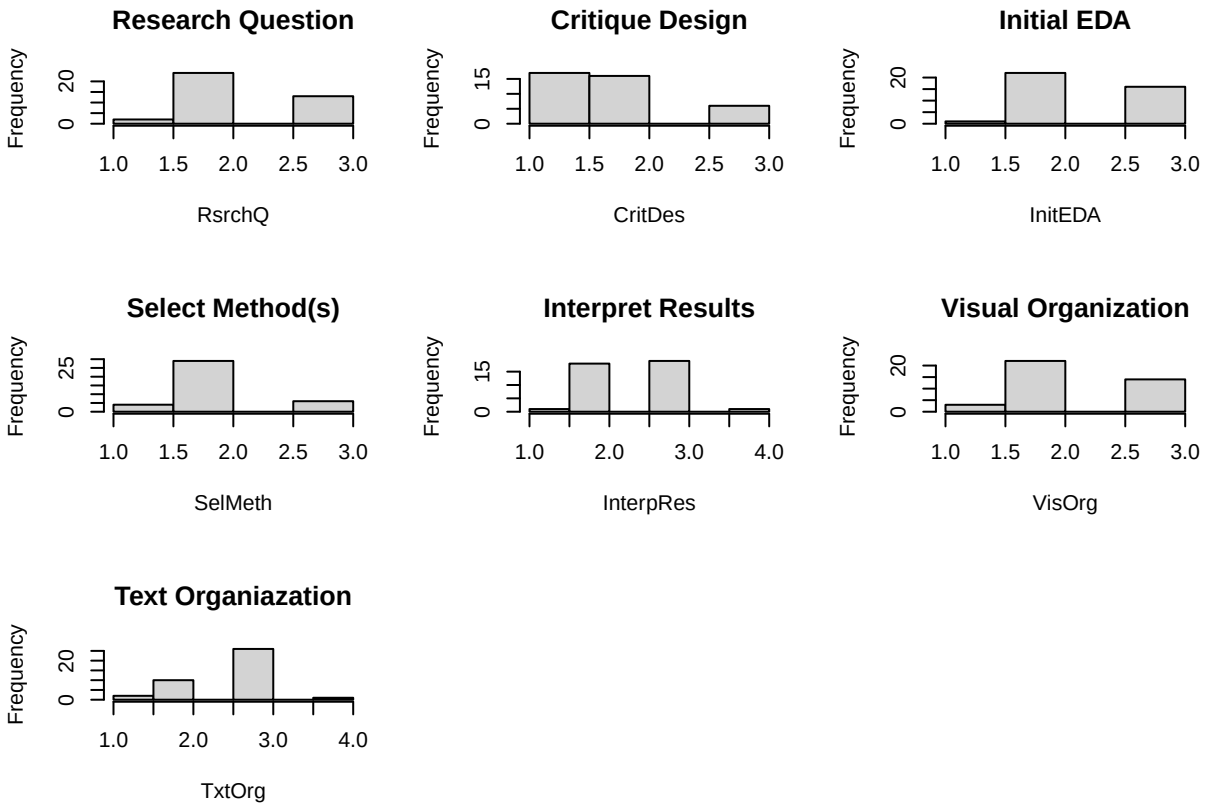
	1 (N=39)	2 (N=39)	3 (N=39)	Total (N=117)
RsrchQ				
Mean (SD)	2.436 (0.641)	2.359 (0.628)	2.256 (0.498)	2.350 (0.592)
CritDes				
Mean (SD)	1.590 (0.715)	2.132 (0.906)	1.897 (0.821)	1.871 (0.840)
InitEDA				
Mean (SD)	2.410 (0.715)	2.564 (0.680)	2.333 (0.701)	2.436 (0.700)
SelMeth				

	1 (N=39)	2 (N=39)	3 (N=39)	Total (N=117)
Mean (SD)	2.128 (0.339)	2.128 (0.469)	1.949 (0.605)	2.068 (0.486)
InterpRes				
Mean (SD)	2.718 (0.456)	2.590 (0.595)	2.154 (0.630)	2.487 (0.610)
VisOrg				
Mean (SD)	2.395 (0.638)	2.641 (0.668)	2.205 (0.656)	2.414 (0.673)
TxtOrg				
Mean (SD)	2.769 (0.583)	2.590 (0.715)	2.436 (0.754)	2.598 (0.696)

Table 1 provides the mean and standard deviation for each rubric of each rater and the overall distribution. The fifth column of the table confirms our observation in the histogram that Critique Design has lower score than the other rubric.(Only one that is lower than 2). Moreover, column 1 to 3 refer to each the statistics for each rater. We can see that rater 1 tends to give lower score on Critique Design (1.59 vs 2.132 vs 1.897). Moreover, rater 3 might be more crucial on rubrics Select methods(1.949 vs 2.128 vs 2.128), and Interpret Result(2.154 vs 2.718 vs 2.590). The rest of the rubrics have the similar patterns for each rater, which are Research question, Initial EDA, Visual Organization, and Text Organization.

Next, we conduct the same analysis on the subset data where only 13 artifacts are counted

```
par(mfrow = c(3,3))
attach(subrating)
hist(RsrchQ, main = "Research Question")
hist(CritDes, main = "Critique Design")
hist(InitEDA, main = "Initial EDA")
hist(SelMeth, main = "Select Method(s)")
hist(InterpRes, main = "Interpret Results")
hist(VisOrg, main = "Visual Organization")
hist(TxtOrg, main = "Text Organiazation")
detach()
```



Compare this plot with the histogram plot we examine with the full data set, we can see that there is no much difference in terms of the score distribution for each rubric. However, for those 13 artifacts, no one has score 4 on Research question, Critique Design, Initial EDA, and visual organization. But overall, it indicates that this can be representative of the whole set of 91 artifacts.

```
table_one1 <- tableby(Rater ~ RsrchQ + CritDes + InitEDA + SelMeth + InterpRes + VisOrg + TxtOrg, data=
```

Table.2 Summary Statistics for 13 Artifacts

	1 (N=13)	2 (N=13)	3 (N=13)	Total (N=39)
RsrchQ				
Mean (SD)	2.385 (0.506)	2.154 (0.689)	2.308 (0.480)	2.282 (0.560)
CritDes				
Mean (SD)	1.615 (0.650)	1.846 (0.801)	1.692 (0.751)	1.718 (0.724)
InitEDA				
Mean (SD)	2.538 (0.660)	2.385 (0.506)	2.231 (0.439)	2.385 (0.544)
SelMeth				
Mean (SD)	2.154 (0.376)	2.077 (0.494)	1.923 (0.641)	2.051 (0.510)
InterpRes				
Mean (SD)	2.615 (0.506)	2.615 (0.650)	2.308 (0.630)	2.513 (0.601)
VisOrg				
Mean (SD)	2.154 (0.555)	2.462 (0.660)	2.231 (0.599)	2.282 (0.605)
TxtOrg				
Mean (SD)	2.769 (0.599)	2.615 (0.650)	2.615 (0.650)	2.667 (0.621)

Similar from last table, after using the 13 artifacts, the Critique Design also has the lowest score among others. For each rater, rater 1 tends to give lowest score on Critique Design than the other two raters. Rater3 also give lower score on Select Method and Visual organization.

Overall, the subset seems like a reasonable representation for the whole set of 91 artifacts.

Appendix: Analysis on differences among raters

Loading tall.csv and focus on 13 artifacts. Changing some variables into factor.

```
tall <- read.csv("tall.csv") %>%
 drop_na() %>%
  filter(Artifact != 5)

tall$Rater <- as.factor(tall$Rater)
tall$Sex <- as.factor(tall$Sex)
tall$Artifact <- as.factor(tall$Artifact)
tall$Rubric <- as.factor(tall$Rubric)

subtall <- tall %>%
  filter(Repeated == 1)
```

To find out whether each raters agree on each other for each, we will check the intraclass correlation(ICC) for each rubric. Generally, a high ICC indicates a high correlation. Fitting lmer() for checking icc. ICC is calculated as va(tau) / (var(tau) + var(sigma)

```
lmer.1 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "RsrchQ"))
lmer.2 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "CritDes"))
lmer.3 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "InitEDA"))
lmer.4 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "SelMeth"))
lmer.5 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "InterpRes"))
lmer.6 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "VisOrg"))
lmer.7 <- lmer(Rating ~ 1 + (1| Artifact), data=subtall %>% filter(Rubric == "TxtOrg"))
```

```
summary(lmer.1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rating ~ 1 + (1 | Artifact)
##    Data: subtall %>% filter(Rubric == "RsrchQ")
##
## 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
```

Reading the var(tau) and var(sigma) from the summary table and calculate the ICC

```
icc1 <- 0.06/(0.06+0.26)
icc1
```

```
## [1] 0.1875
```

We illustrate one example as the example of ICC calculation, the rest of calculation for ICC is conducted in the similar way. Below is the table for the ICC score for each rubric using the subset for the 13 artifacts.

Table 3. ICC Statistics with 13 Artifacts

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
ICC	0.188	0.574	0.500	0.519	0.222	0.594	0.145

Table 3 shows the ICC's for each rubric using subset data. for each column, the top row indicates the name of the rubric. The second row is the respective ICC. From the table, we can see that Critique Design, Initial EDA, Select Methods, and Visual Organization have ICC over 0.5, which indicate a medium agreement among raters. However, Research question, Interpret Results, and Text Organization have lower ICC that are below 0.5. It suggests that raters in those rubrics tend to have low agreement.

Next, we run the same thing the the whole data set

```
lmer.8 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "RsrchQ"))
lmer.9 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "CritDes"))
lmer.10 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "InitEDA"))
lmer.11 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "SelMeth"))
lmer.12 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "InterpRes"))
lmer.13 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "VisOrg"))
lmer.14 <- lmer(Rating ~ 1 + (1| Artifact), data=tall %>% filter(Rubric == "TxtOrg"))
```

Using the same method, we get the ICC with full dataset indicated in the table below.

Table 4. ICC statistics with Full dataset

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
ICC	0.210	0.673	0.687	0.472	0.220	0.661	0.188

Table 4 shows the ICC's for each rubric for the whole dataset tall. We can see that it has the similar pattern as what shows in last table. Research question, Interpret Results, Selected Method, and Text Organization have ICC's lower than 0.5, which suggests that raters on those rubric have low agreement. Critique Design, Initial EDA, visual Organization have ICC's higher between 0.5 and 0.75. It suggest that raters have a

medium agreement on those rubric. The only difference from these two tables is that Select Methods change from 0.519 to 0.477. But we think this difference is reasonable after switching to bigger data set.

This disagreement might be an indicator of random effect on the rater variable for further model selection

Next, we makes two way tables for examine which rater differs from others. To do this, we make 2-way table for counts for the rating of each pairs of raters, on each rubric. For each table, we calculate the percentage of observations on the main diagonal. This can be seen as the percent exact agreement between two raters.

```
rating13 <- rating %>%
  filter(Repeated==1)
```

```
for (i in names(rating13)[7:13]){
  f1rq <- factor(rating13[rating13$Rater==1,as.character(i)],levels=1:4)
  f2rq <- factor(rating13[rating13$Rater==2,as.character(i)],levels=1:4)
  f3rq <- factor(rating13[rating13$Rater==3,as.character(i)],levels=1:4)

  print(i)

  table1 <- table(f1rq, f2rq)
  print(table1)
  print(sum(diag(table1)) / sum(rowSums(table1)))

  table2 <- table(f1rq,f3rq)
  print(table2)
  print(sum(diag(table2)) / sum(rowSums(table2)))

  table3 <- table(f2rq, f3rq)
  print(table3)
  print(sum(diag(table3)) / sum(rowSums(table3)))

  print(" ")
}
```

```
## [1] "RsrchQ"
##     f2rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 1 4 3 0
##    3 1 3 1 0
##    4 0 0 0 0
## [1] 0.3846154
##     f3rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 0 7 1 0
##    3 0 2 3 0
##    4 0 0 0 0
## [1] 0.7692308
##     f3rq
## f2rq 1 2 3 4
##    1 0 2 0 0
##    2 0 5 2 0
##    3 0 2 2 0
```

```
##    4 0 0 0 0
## [1] 0.5384615
## [1] " "
## [1] "CritDes"
##     f2rq
## f1rq 1 2 3 4
##    1 3 2 1 0
##    2 2 3 1 0
##    3 0 0 1 0
##    4 0 0 0 0
## [1] 0.5384615
##     f3rq
## f1rq 1 2 3 4
##    1 4 2 0 0
##    2 2 3 1 0
##    3 0 0 1 0
##    4 0 0 0 0
## [1] 0.6153846
##     f3rq
## f2rq 1 2 3 4
##    1 5 0 0 0
##    2 1 3 1 0
##    3 0 2 1 0
##    4 0 0 0 0
## [1] 0.6923077
## [1] " "
## [1] "InitEDA"
##     f2rq
## f1rq 1 2 3 4
##    1 0 1 0 0
##    2 0 4 0 0
##    3 0 3 5 0
##    4 0 0 0 0
## [1] 0.6923077
##     f3rq
## f1rq 1 2 3 4
##    1 0 1 0 0
##    2 0 4 0 0
##    3 0 5 3 0
##    4 0 0 0 0
## [1] 0.5384615
##     f3rq
## f2rq 1 2 3 4
##    1 0 0 0 0
##    2 0 8 0 0
##    3 0 2 3 0
##    4 0 0 0 0
## [1] 0.8461538
## [1] " "
## [1] "SelMeth"
##     f2rq
## f1rq  1  2  3  4
##    1  0  0  0  0
##    2  1 10  0  0
```

```
##    3  0  0  2  0
##    4  0  0  0  0
## [1] 0.9230769
##     f3rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 3 7 1 0
##    3 0 1 1 0
##    4 0 0 0 0
## [1] 0.6153846
##     f3rq
## f2rq 1 2 3 4
##    1 1 0 0 0
##    2 2 7 1 0
##    3 0 1 1 0
##    4 0 0 0 0
## [1] 0.6923077
## [1] " "
## [1] "InterpRes"
##     f2rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 0 3 1 1
##    3 0 3 5 0
##    4 0 0 0 0
## [1] 0.6153846
##     f3rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 1 3 1 0
##    3 0 4 4 0
##    4 0 0 0 0
## [1] 0.5384615
##     f3rq
## f2rq 1 2 3 4
##    1 0 0 0 0
##    2 1 4 1 0
##    3 0 2 4 0
##    4 0 1 0 0
## [1] 0.6153846
## [1] " "
## [1] "VisOrg"
##     f2rq
## f1rq 1 2 3 4
##    1 1 0 0 0
##    2 0 4 5 0
##    3 0 1 2 0
##    4 0 0 0 0
## [1] 0.5384615
##     f3rq
## f1rq 1 2 3 4
##    1 1 0 0 0
##    2 0 7 2 0
##    3 0 1 2 0
```

```
##    4 0 0 0 0
## [1] 0.7692308
##     f3rq
## f2rq 1 2 3 4
##    1 1 0 0 0
##    2 0 5 0 0
##    3 0 3 4 0
##    4 0 0 0 0
## [1] 0.7692308
## [1] " "
## [1] "TxtOrg"
##     f2rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 0 2 2 0
##    3 0 1 7 0
##    4 1 0 0 0
## [1] 0.6923077
##     f3rq
## f1rq 1 2 3 4
##    1 0 0 0 0
##    2 1 1 2 0
##    3 0 1 7 0
##    4 0 1 0 0
## [1] 0.6153846
##     f3rq
## f2rq 1 2 3 4
##    1 0 1 0 0
##    2 1 0 2 0
##    3 0 2 7 0
##    4 0 0 0 0
## [1] 0.5384615
## [1] " "
```

we calculate the agreement rates for each pair of rater in each rubric. Table showns as below

Table 5. Percent Agreement with Full dataset

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
rater1 vs rater2	38.5%	53.8%	69,2%	74.4%	61.5%	53.8%	69.2%
rater1 vs rater3	76.9%	61.5%	53.8%	56.4%	53.8%	76.9%	61.5%
rater2 vs rater3	53.8%	69.2%	84.6%	43.5%	61.5%	76.9%	53.8%

Appendix: Regression Analysis

Here we consider the multilevel model since each artifact differs from each other in terms of each rubric, which can be seen as the cluster. On the other hand, from the former analyses, we notice that the rubric and rater can potentially bring the random effects as well. So we start by considering fixed effects to the following models. Then we add potential fixed effects and random effects.

```
lmer.fit0 <- lmer(Rating ~ (1 + Rubric | Artifact), data=tall,REML = FALSE)
```

```
lmer.fit1 <- lmer(Rating ~(0 + Rubric | Artifact), data=tall,REML = FALSE)
```

```
anova(lmer.fit0, lmer.fit1)
```

```
## Data: tall
## Models:
## lmer.fit0: Rating ~ (1 + Rubric | Artifact)
## lmer.fit1: Rating ~ (0 + Rubric | Artifact)
##           npar  AIC  BIC  logLik deviance Chisq Df Pr(>Chisq)
## lmer.fit0   30 1527 1668 -733.52     1467
## lmer.fit1   30 1527 1668 -733.52     1467 0.003  0  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding random intercept does not improve the model, we will go on with lmer.fit1

```
summary(lmer.fit1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ (0 + Rubric | Artifact)
##    Data: tall
##
##      AIC      BIC   logLik deviance df.resid
##   1527.0   1668.0   -733.5   1467.0      780
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0167 -0.4921 -0.0813  0.5249  3.7859
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.63880  0.7993
##           RubricInitEDA   0.38190  0.6180    0.25
##           RubricInterpRes 0.25549  0.5055   -0.01  0.78
##           RubricRsrchQ    0.17264  0.4155    0.38  0.50  0.74
##           RubricSelMeth   0.09484  0.3080    0.56  0.36  0.40  0.25
##           RubricTxtOrg    0.40336  0.6351    0.02  0.69  0.80  0.64  0.23
##           RubricVisOrg    0.31791  0.5638    0.17  0.78  0.77  0.59  0.28  0.79
##  Residual                 0.19449  0.4410
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  2.23158    0.03989   55.94
```

Next, we add fixed effect Rubric

```
lmer.fit2 <- lmer(Rating ~ 1 + Rubric + (0 + Rubric | Artifact), data=tall,REML = FALSE)
```

We then compare lmer.fit1 and lmer.fit2

```
anova(lmer.fit1, lmer.fit2)
```

```
## Data: tall
## Models:
## lmer.fit1: Rating ~ (0 + Rubric | Artifact)
## lmer.fit2: Rating ~ 1 + Rubric + (0 + Rubric | Artifact)
##           npar    AIC  BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit1   30 1527.0 1668 -733.52   1467.0
## lmer.fit2   36 1474.9 1644 -701.43   1402.9 64.181  6   6.34e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both AIC and BIC prefer lmer.fit2.

Next, we add rater as fixed effect

```
lmer.fit3 <- lmer(Rating ~ 1 + Rubric + Rater + (0 + Rubric | Artifact), data=tall,REML = FALSE)
```

```
anova(lmer.fit2, lmer.fit3)
```

```
## Data: tall
## Models:
## lmer.fit2: Rating ~ 1 + Rubric + (0 + Rubric | Artifact)
## lmer.fit3: Rating ~ 1 + Rubric + Rater + (0 + Rubric | Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit2   36 1474.9 1644.0 -701.43   1402.9
## lmer.fit3   38 1466.0 1644.5 -694.99   1390.0 12.889  2   0.001589 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both AIC and BIC prefer lmer.fit3

```
lmer.fit4 <- lmer(Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact), data=tall,REML = FALSE)
```

```
anova(lmer.fit3,lmer.fit4)
```

```
## Data: tall
## Models:
## lmer.fit3: Rating ~ 1 + Rubric + Rater + (0 + Rubric | Artifact)
## lmer.fit4: Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit3   38 1466.0 1644.5 -694.99   1390.0
## lmer.fit4   39 1467.5 1650.7 -694.74   1389.5 0.4925  1     0.4828
```

AIC and BIC both slightly prefer lmer.fit4.

```
lmer.fit5 <- lmer(Rating ~ 1 + Repeated + Sex + Rubric + Rater + (0 + Rubric | Artifact), data=tall,REM
```

```
anova(lmer.fit4,lmer.fit5)
```

```
## Data: tall
## Models:
## lmer.fit4: Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact)
## lmer.fit5: Rating ~ 1 + Repeated + Sex + Rubric + Rater + (0 + Rubric |
## lmer.fit5:     Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit4   39 1467.5 1650.7 -694.74   1389.5
## lmer.fit5   40 1469.2 1657.1 -694.59   1389.2 0.3017  1     0.5828
```

Here AIC and BIC prefer lmer.fit4. Next we add Semester

```
lmer.fits <- lmer(Rating ~ 1 + Semester + Sex + Rubric + Rater + (0 + Rubric | Artifact), data=tall,REM
```

```
anova(lmer.fit4, lmer.fits)
```

```
## Data: tall
## Models:
## lmer.fit4: Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact)
## lmer.fits: Rating ~ 1 + Semester + Sex + Rubric + Rater + (0 + Rubric |
## lmer.fits:     Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit4   39 1467.5 1650.7 -694.74   1389.5
## lmer.fits   40 1466.0 1653.9 -693.01   1386.0 3.4567  1      0.063 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even though the AIC slightly prefer lmer.fits, the BIC still prefer lmer.fit4. Here, we still consider fiexed effect shown as lmer.fit4

Next we consider interaction term. Since we are interested to find out more about the effect of sex, we consider sex*rubric

```
lmer.fit6 <- lmer(Rating ~ 1 + Sex*Rubric + Rater + (0 + Rubric | Artifact), data=tall,REML = FALSE)
```

```
anova(lmer.fit4,lmer.fit6)
```

```
## Data: tall
## Models:
## lmer.fit4: Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact)
## lmer.fit6: Rating ~ 1 + Sex * Rubric + Rater + (0 + Rubric | Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit4   39 1467.5 1650.7 -694.74   1389.5
## lmer.fit6   45 1472.5 1683.9 -691.26   1382.5 6.9505  6     0.3254
```

The AIC and BIC still prefer lmer.fit4. Next, we consider rubric*Rater.

```
lmer.fit7 <- lmer(Rating ~ 1 + Sex + Rubric + Rater + Rubric*Rater + (0 + Rubric | Artifact), data=tall
```

```
anova(lmer.fit4, lmer.fit7)
```

```
## Data: tall
## Models:
## lmer.fit4: Rating ~ 1 + Sex + Rubric + Rater + (0 + Rubric | Artifact)
## lmer.fit7: Rating ~ 1 + Sex + Rubric + Rater + Rubric * Rater + (0 + Rubric |
## lmer.fit7:     Artifact)
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmer.fit4   39 1467.5 1650.7 -694.74   1389.5
## lmer.fit7   51 1458.3 1697.8 -678.13   1356.3 33.212 12  0.0008971 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AIC and BIC prefer lmer.fit7. We can see that AIC for lmer.fit7 is almost 10 lower than AIC for lmer.fit4. Thus, it's worth considering the interaction term.

Next, we consider adding random effect. we use the fitLMER.fnc() function to conduct this step. As we discuss above, both Rubric and Rater can bring the random effect. Thus, we will consider them in the following steps.

We consider rater and sex and semester.

```
library(LMERConvenienceFunctions)
lmer.fit8 <- fitLMER.fnc(lmer.fit7, ran.effects=c("(1|Rater)", "(1|Sex)", "(1|Semester)"), method="AIC"
```

```
## ========================================================
## ===              backfitting fixed effects         ===
## ========================================================
## setting REML to FALSE
## processing model terms of interaction level 2
##   all terms of interaction level 2 significant
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## pruning random effects structure ...
##   nothing to prune
## ========================================================
## ===            forwardfitting random effects       ===
## ========================================================
## evaluating addition of (1|Rater) to model
##  log-likelihood ratio test p-value = 1
##  not adding (1|Rater) to model
## evaluating addition of (1|Sex) to model
##  log-likelihood ratio test p-value = 0.9994819
##  not adding (1|Sex) to model
## evaluating addition of (1|Semester) to model
##  log-likelihood ratio test p-value = 0.6674219
##  not adding (1|Semester) to model
## ========================================================
## ===            re-backfitting fixed effects        ===
## ========================================================
## setting REML to FALSE
```

```
## processing model terms of interaction level 2
##   all terms of interaction level 2 significant
## processing model terms of interaction level 1
##   all terms of interaction level 1 significant
## resetting REML to TRUE
## pruning random effects structure ...
##   nothing to prune
## log file is C:\Users\EugLu\AppData\Local\Temp\Rtmp4uQkXh/fitLMER_log_Sat_Dec_12_20-54-24_2020.txt
```

We see that fitLMER.fnc() does not prefer other random effect. Here, we pick model 7 as our final model

```
summary(lmer.fit7)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Sex + Rubric + Rater + Rubric * Rater + (0 + Rubric |
##     Artifact)
##    Data: tall
##
##      AIC      BIC   logLik deviance df.resid
##   1458.3   1697.8   -678.1   1356.3      759
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9239 -0.5217 -0.0393  0.4916  3.6463
##
## Random effects:
##  Groups   Name            Variance Std.Dev. Corr
##  Artifact RubricCritDes   0.49054  0.7004
##           RubricInitEDA   0.34174  0.5846   0.44
##           RubricInterpRes 0.14808  0.3848   0.35 0.81
##           RubricRsrchQ    0.16353  0.4044   0.62 0.42 0.69
##           RubricSelMeth   0.07636  0.2763   0.42 0.59 0.75 0.35
##           RubricTxtOrg    0.26350  0.5133   0.42 0.63 0.68 0.54 0.66
##           RubricVisOrg    0.26175  0.5116   0.34 0.71 0.69 0.51 0.43 0.78
##  Residual                 0.17981  0.4240
## Number of obs: 810, groups:  Artifact, 90
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)              1.67909    0.11827  14.197
## SexM                     0.05680    0.07667   0.741
## RubricInitEDA            0.74551    0.13515   5.516
## RubricInterpRes          1.01335    0.13321   7.607
## RubricRsrchQ             0.74831    0.12267   6.100
## RubricSelMeth            0.42646    0.12876   3.312
## RubricTxtOrg             1.04839    0.13395   7.827
## RubricVisOrg             0.68256    0.13792   4.949
## Rater2                   0.36612    0.13117   2.791
## Rater3                   0.21363    0.13119   1.628
## RubricInitEDA:Rater2    -0.30755    0.17025  -1.807
## RubricInterpRes:Rater2  -0.53491    0.16787  -3.186
## RubricRsrchQ:Rater2     -0.49833    0.15924  -3.129
## RubricSelMeth:Rater2    -0.39568    0.16275  -2.431
```

```
## RubricTxtOrg:Rater2    -0.58384    0.16924  -3.450
## RubricVisOrg:Rater2    -0.14446    0.17229  -0.839
## RubricInitEDA:Rater3   -0.29581    0.17057  -1.734
## RubricInterpRes:Rater3 -0.75051    0.16828  -4.460
## RubricRsrchQ:Rater3    -0.37090    0.15953  -2.325
## RubricSelMeth:Rater3   -0.40852    0.16312  -2.504
## RubricTxtOrg:Rater3    -0.48382    0.16960  -2.853
## RubricVisOrg:Rater3    -0.32945    0.17266  -1.908
```

```
#critides

0.49054/(0.49054+0.17981)
```

```
## [1] 0.731767
```

```
#InitEDA
0.34174/ (0.34174+0.17981)
```

```
## [1] 0.6552392
```

```
#InterpRes
0.14808/ (0.14808+0.17981)
```

```
## [1] 0.4516149
```

```
#ReschQ
0.16353/ (0.16353+0.17981)
```

```
## [1] 0.4762917
```

```
#Selmeth
0.07636  / (0.07636  +0.17981)
```

```
## [1] 0.2980833
```

```
#TxtOrg
0.26350  / (0.26350  +0.17981)
```

```
## [1] 0.5943922
```

```
#Visorg
0.68256  / (0.68256  +0.17981)
```

```
## [1] 0.7914932
```

Table 3. ICC Statistics in final models

	RsrchQ	CritDes	InitEDA	SelMeth	InterpRes	VisOrg	TxtOrg
ICC	0.476	0.732	0.655	0.298	0.452	0.791	0.594

Last, The vif does not suggest any muticollinearity

```
library(car)
vif(lmer.fit7)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Sex                1.001845  1        1.000922
## Rubric           203.198739  6        1.557137
## Rater             35.610902  2        2.442844
## Rubric:Rater    4242.168927 12        1.416281
```

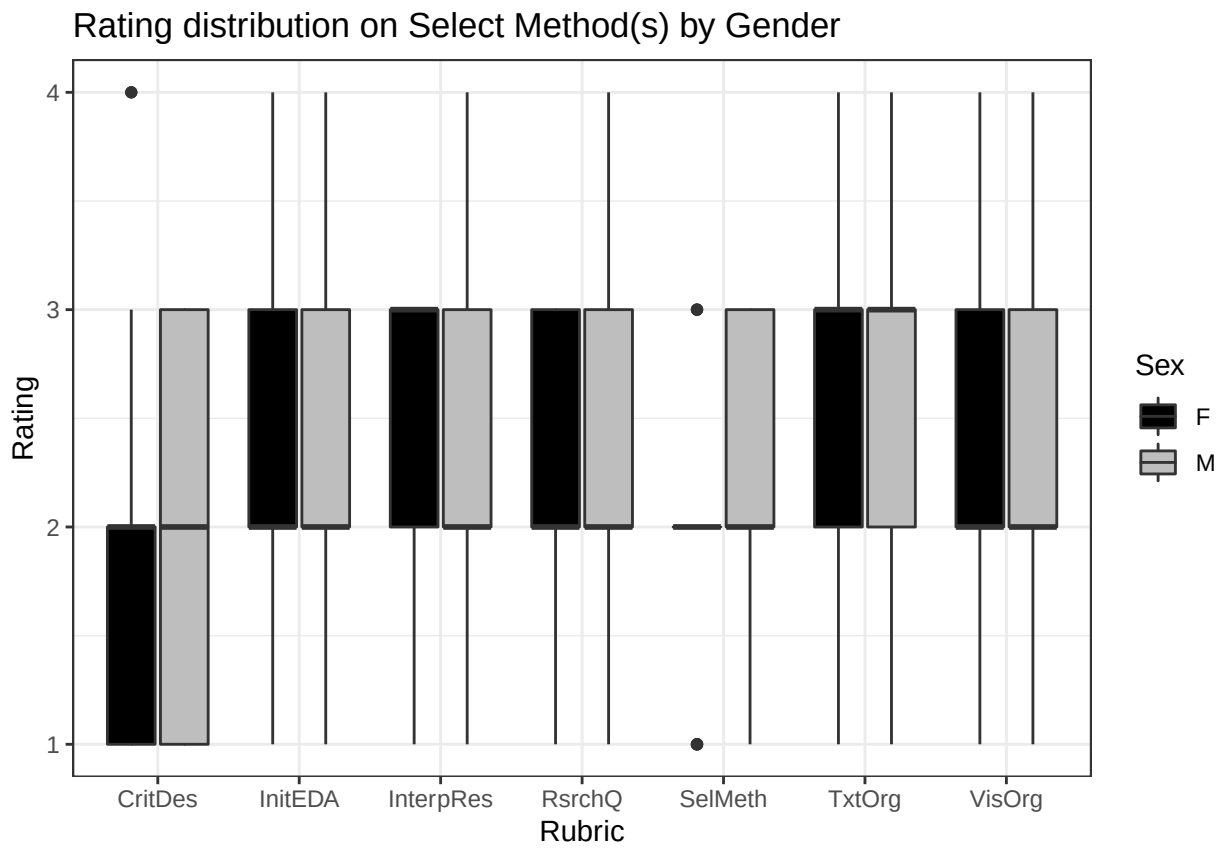
Appendix: Additional information about the data

```
table(rating$Sex)
```

```
##
## --  F  M
##  1 64 52
```

If appears that the number of female students is greater than male students. We might compare the gender distribution on each rubric

```
ggplot(data = tall, aes(x= Rubric, y=Rating, fill= Sex))+
  geom_boxplot()+
  scale_fill_manual(values=c("black", "grey")) +
  labs(title="Rating distribution on Select Method(s) by Gender")+
  theme_bw()
```



It looks like the only difference between male and female student on the rating is select methods. Next, we take look in to Selmeth

```
sel <- tall %>%
  filter(Rubric == "SelMeth") %>%
  group_by(Sex) %>%
  summarise(count = n(),
            average = round(mean(Rating),3),
            median = round(median(Rating),3),
            SD = round(sd(Rating),3))
```

```
knitr::kable(sel)
```

Sex	count	average	median	SD
F	64	1.984	2	0.333
M	52	2.154	2	0.607

Summary statistics for critique design

```
sel2 <- tall %>%
  filter(Rubric == "CritDes") %>%
  group_by(Sex) %>%
  summarise(count = n(),
            average = round(mean(Rating),3),
```

```
            median = round(median(Rating),3),
            SD = round(sd(Rating),3))
```

```
knitr::kable(sel2)
```

Sex	count	average	median	SD
F	63	1.778	2	0.888
M	52	1.962	2	0.766