

The impact of double blind reviewing at EvoLang 12: statistics

Contents

| | |
|--|-----------|
| Introduction | 1 |
| Data | 1 |
| Loading data for first analysis | 1 |
| Plots | 4 |
| Review ranks by gender and student status | 13 |
| Mixed effects model | 15 |
| Permutation test | 19 |
| Decision tree exploration | 22 |
| Readability scores | 23 |
| Reading scores and review scores | 46 |
| Influence of last author | 52 |

Introduction

Data

This script uses the data file `EvoLang_Scores_8_to_12.csv`:

- `conference`: Which conference the paper was submitted to
- `gender`: Gender of first author
- `Score.Mean`: Mean raw score given by reviewers (scaled between 0 and 1, higher = better paper)
- `student`: The student status of the first author at submission.

All variables with an underscore are measures of readability. Below we calculate a variable `review`, which represents the type of review (Single / Double blind).

Loading data for first analysis

Load libraries.

```
# Load data
library(lattice)
library(ggplot2)
library(gplots)
library(lme4)
library(magrittr)
library(qwraps2)
library(car)
```

```

library(caret)
library(dplyr)
library(party)
library(lmerTest)
library(stargazer)

# read data
allData = read.csv("../data/EvoLang_Scores_8_to_12.csv", stringsAsFactors = F)
# relabel factor
allData$FirstAuthorGender = factor(allData$FirstAuthorGender, labels=c("F", "M"))
allData$review = factor(c("Single", "Double")[(allData$conference %in% c("E11", "E12"))+1])
allData$conference = factor(allData$conference, levels = c("E8", "E9", "E10", "E11", "E12"))
allData$format = factor(allData$format)

allData$student[!is.na(allData$student) &
  allData$student=="Faculty"] = "Non-Student"
allData$student[!is.na(allData$student) &
  allData$student=="EC"] = "Non-Student"
allData$student = factor(allData$student)

#allData$Score.mean = scale(allData$Score.mean)

for(conf in levels(allData$conference)){
  allData$Score.mean[allData$conference==conf] = scale(allData$Score.mean[allData$conference==conf])
}

```

Look at the distribution of submissions:

```

table(allData$FirstAuthorGender, allData$conference)

##
##      E8  E9 E10 E11 E12
##  F   58  52  67  76  84
##  M   94 130 124 119 122

prop.table(table(allData$FirstAuthorGender, allData$conference), 2)

##
##      E8      E9      E10      E11      E12
##  F 0.3815789 0.2857143 0.3507853 0.3897436 0.4077670
##  M 0.6184211 0.7142857 0.6492147 0.6102564 0.5922330

gtable = table(allData$FirstAuthorGender, allData$conference, allData$student)
write.csv(cbind(t(gtable[, , 1]), t(gtable[, , 2])),
  "../results/CountTable.csv")
gtable

## , , = Non-Student
##
##
##      E8 E9 E10 E11 E12
##  F  0 34  55  41  54
##  M  0 85  94  77  93
##
## , , = Student
##

```

```
##
##      E8 E9 E10 E11 E12
##  F   0 18  12  35  30
##  M   0 45  30  42  29
```

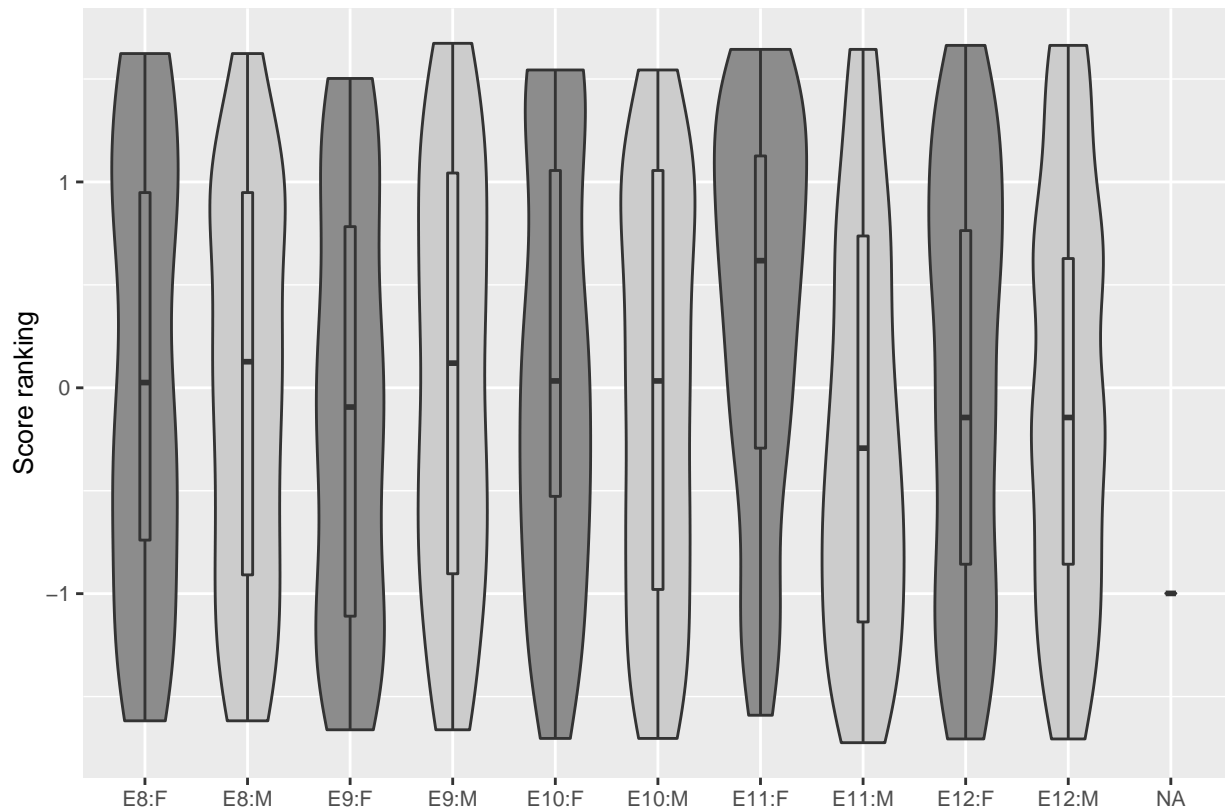
Plots

Rank by gender. It seems that the difference in E11 is not replicated in E12.

```
source("../analysis/summarySE.r")
p2 <- ggplot(allData,
             aes((conference):(FirstAuthorGender), Score.mean,
                 fill=FirstAuthorGender))

p2 <- p2 + geom_violin() + geom_boxplot(width=0.1) +
  theme(text=element_text(size=20), legend.position="none") +
  scale_y_continuous(name="Score ranking")+
  scale_x_discrete(name="")+
  scale_fill_grey(start = 0.55, end=0.8) +
  theme(text = element_text(size=10))
```

p2



```
pdf("../results/Results_Gender_3conf.pdf", width = 12, height= 6)
p2
dev.off()
```

```
## pdf
## 2
```

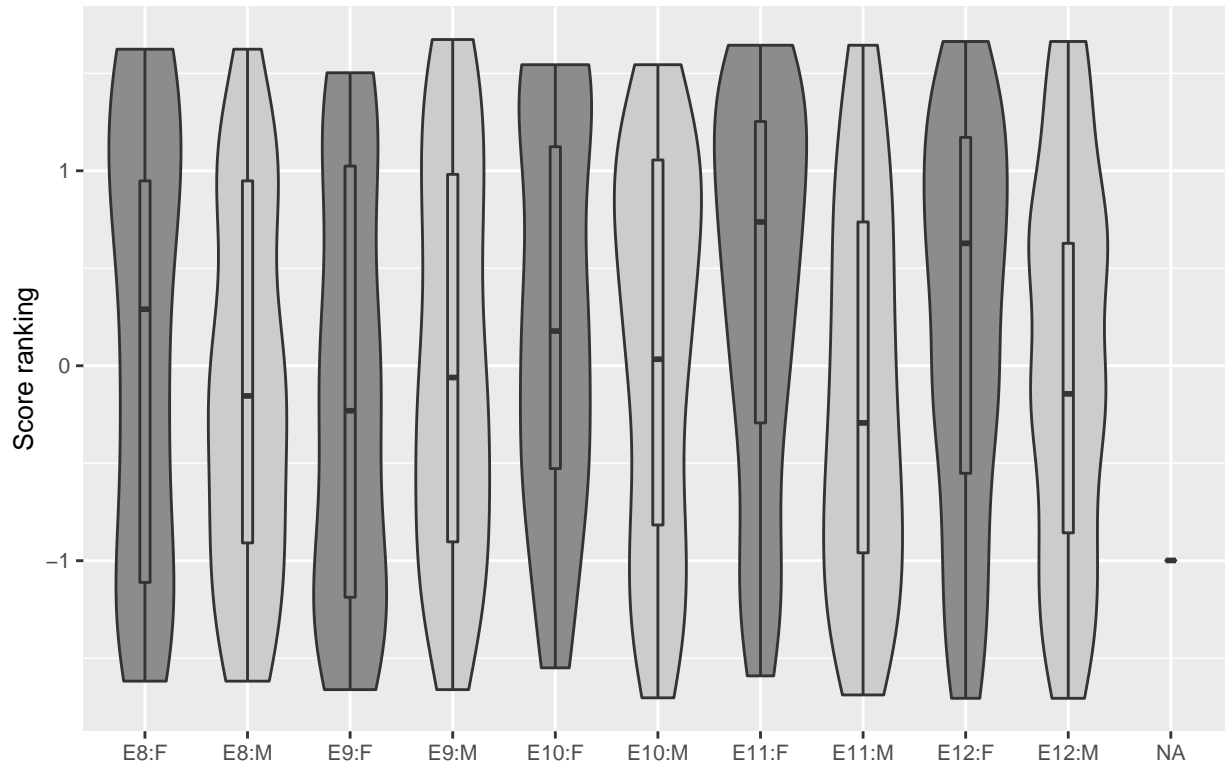
```
p2Abstract <- ggplot(allData[allData$format=="Abstract",],
                    aes((conference):(FirstAuthorGender), Score.mean,
                        fill=FirstAuthorGender))
```

```

p2Abstract <- p2Abstract + geom_violin() + geom_boxplot(width=0.1) +
  theme(text=element_text(size=20), legend.position="none") +
  scale_y_continuous(name="Score ranking")+
  scale_x_discrete(name="")+
  scale_fill_grey(start = 0.55, end=0.8) +
  theme(text = element_text(size=10)) +
  ggtitle("Scores for abstracts only")
p2Abstract

```

Scores for abstracts only

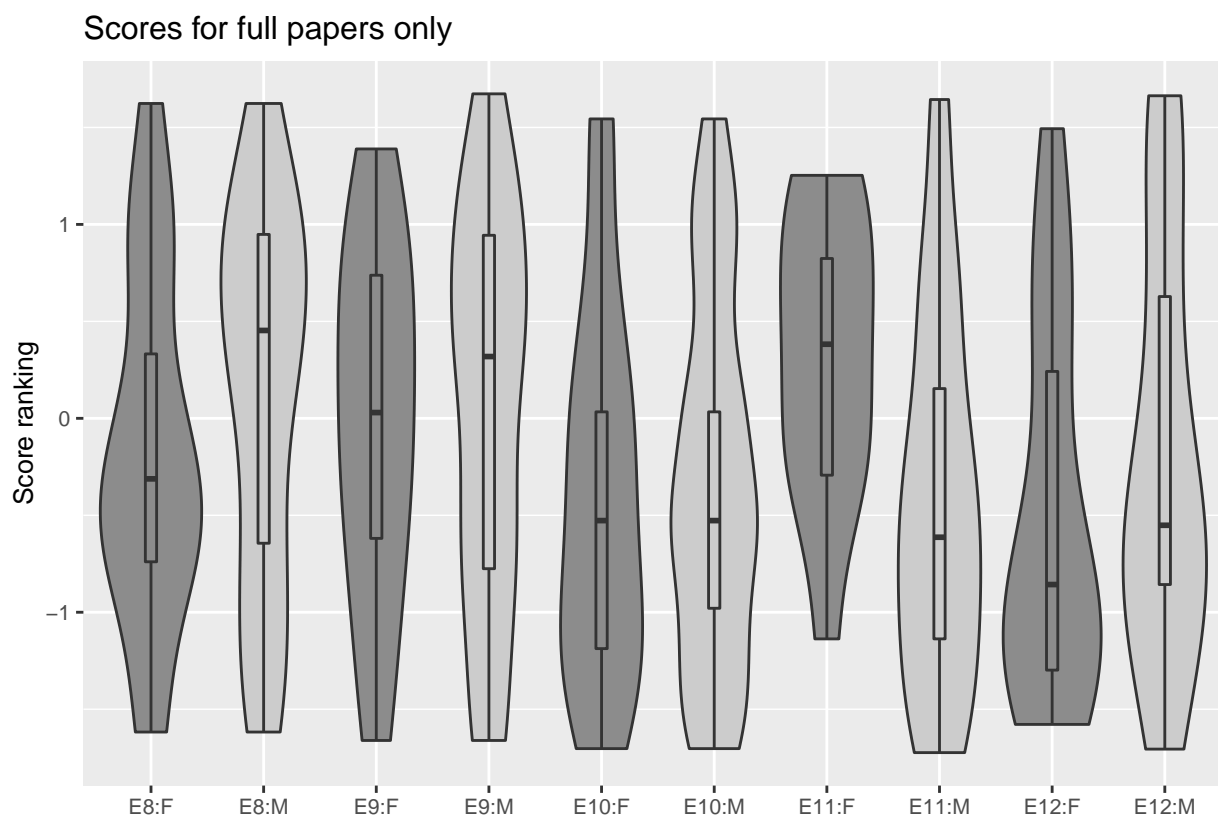


```

p2Paper <- ggplot(allData[allData$format=="Paper",],
  aes((conference):(FirstAuthorGender), Score.mean,
    fill=FirstAuthorGender))

p2Paper <- p2Paper + geom_violin() + geom_boxplot(width=0.1) +
  theme(text=element_text(size=20), legend.position="none") +
  scale_y_continuous(name="Score ranking")+
  scale_x_discrete(name="")+
  scale_fill_grey(start = 0.55, end=0.8) +
  theme(text = element_text(size=10)) +
  ggtitle("Scores for full papers only")
p2Paper

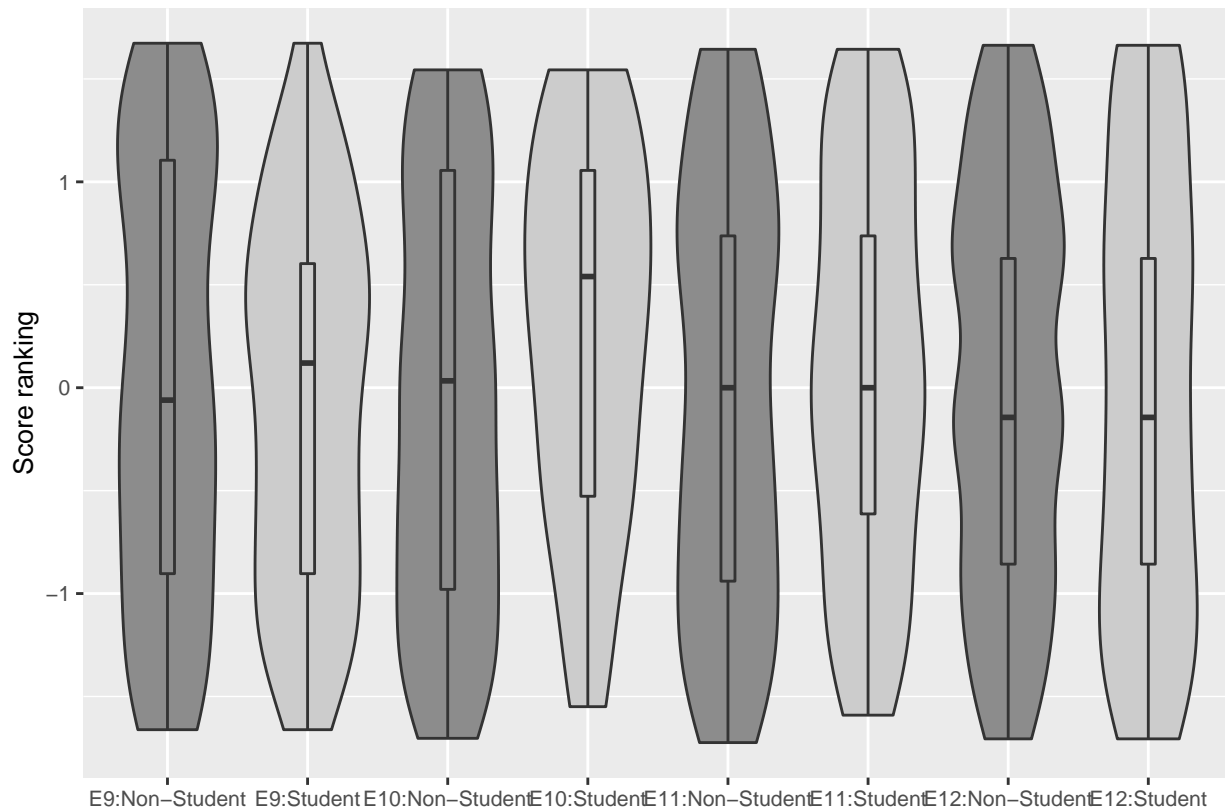
```



Rank by student status in each conference.

```
p <- ggplot(allData[complete.cases(allData),], aes(conference:student, Score.mean, fill=student))

p <- p + geom_violin() + geom_boxplot(width=0.1) +
  theme(text=element_text(size=20), legend.position="none") +
  scale_y_continuous(name="Score ranking")+
  scale_x_discrete(name="")+
  scale_fill_grey(start = 0.55, end=0.8)+
  theme(text = element_text(size=10))
p
```



```
pdf("../results/Results_Student_3conf.pdf", width = 12, height= 6)
```

```
p
dev.off()
```

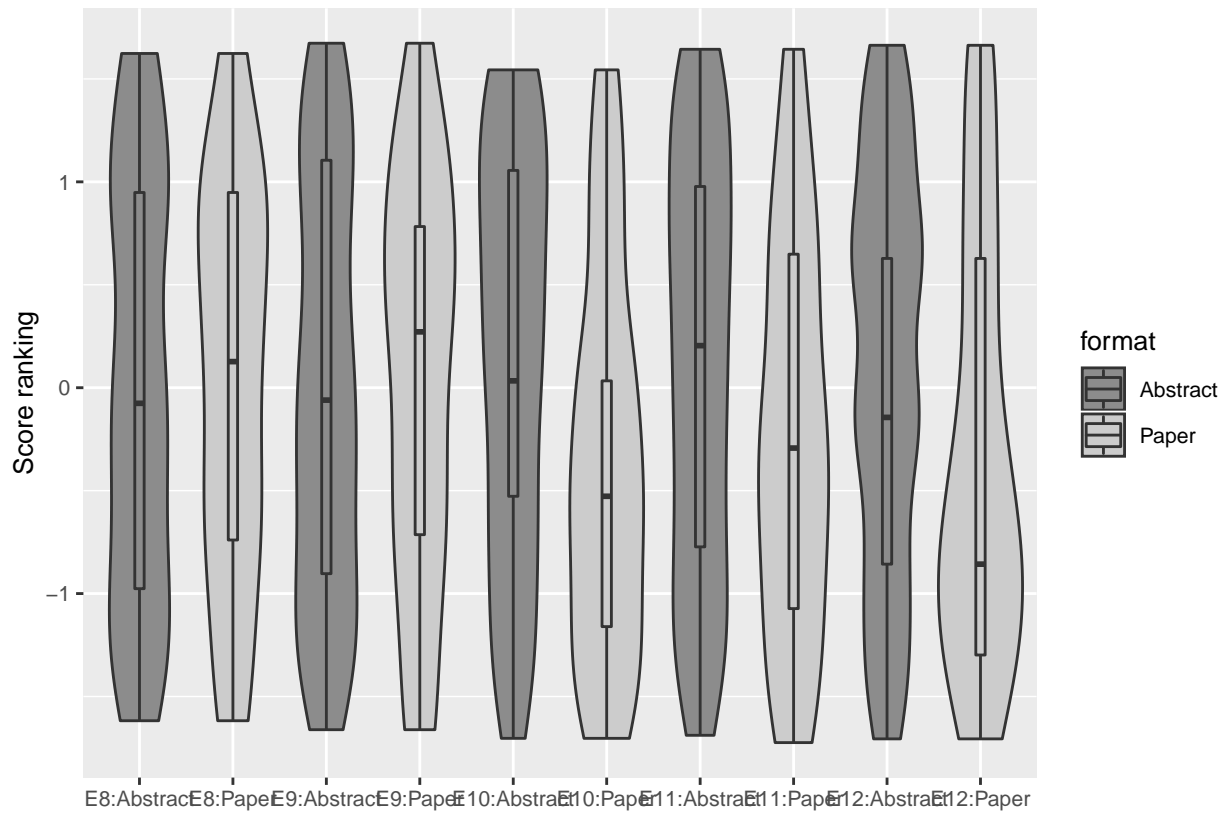
```
## pdf
## 2
```

Format:

```
p <- ggplot(allData, aes(conference:format, Score.mean, fill=format))
```

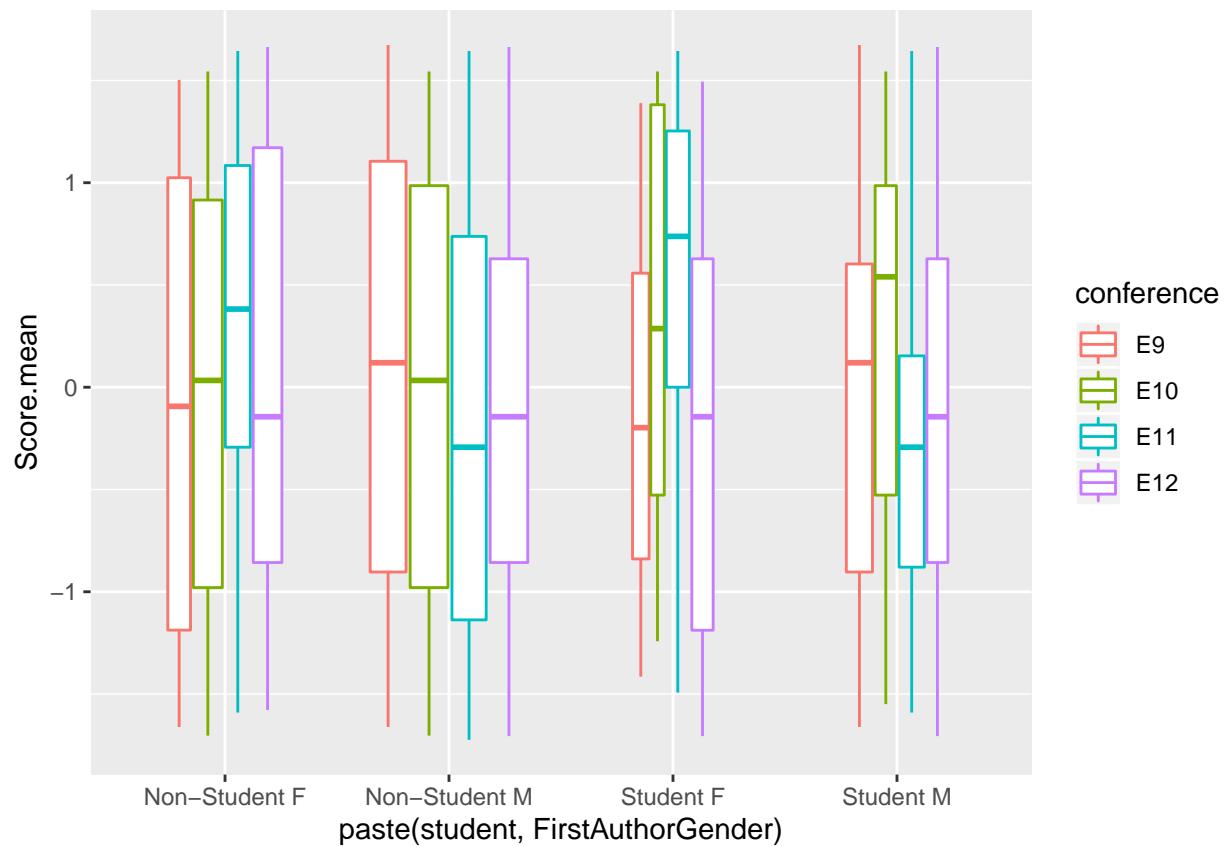
```
p <- p + geom_violin() + geom_boxplot(width=0.1) +
  theme(text=element_text(size=10)) +
  scale_y_continuous(name="Score ranking")+
  scale_x_discrete(name="")+
  scale_fill_grey(start = 0.55, end=0.8)
```

```
p
```



Combined student and gender:

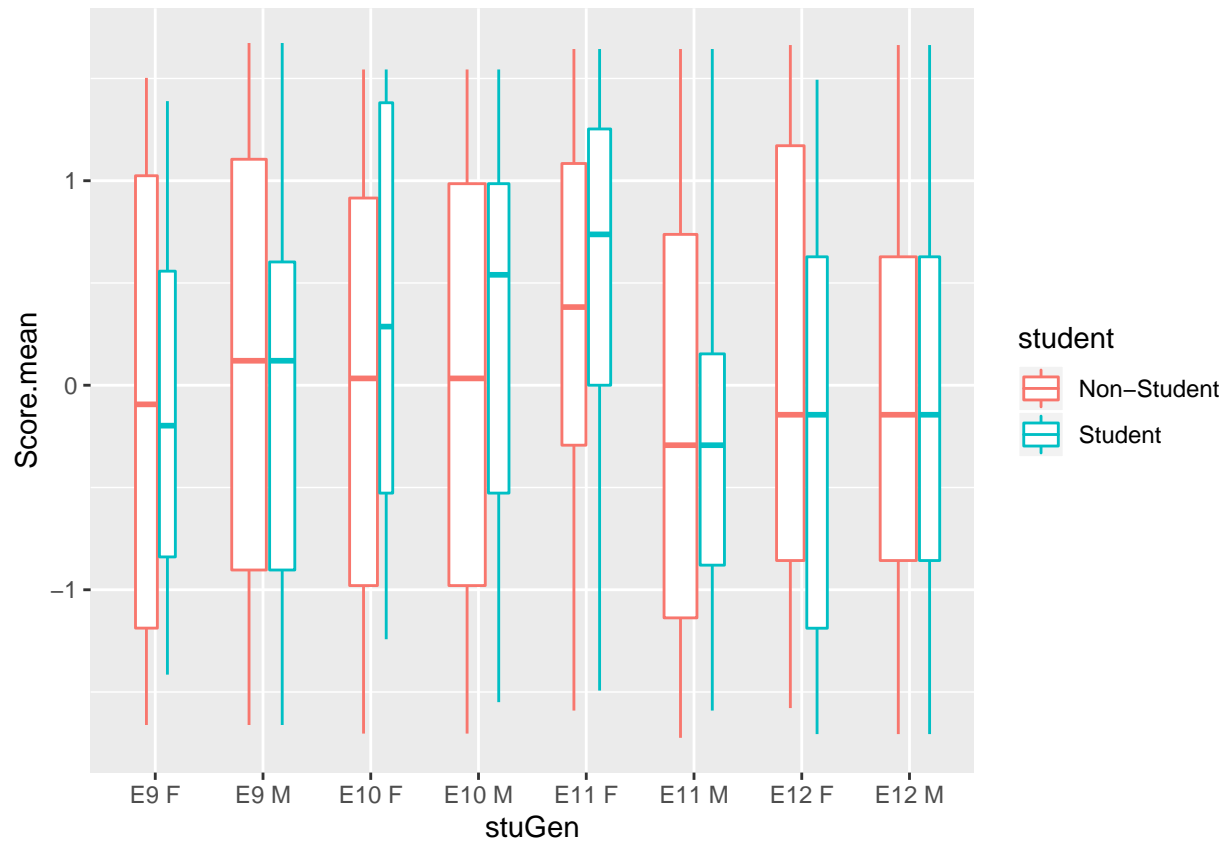
```
ggplot(allData[allData$conference!="E8",],
  aes(y=Score.mean,x=paste(student,FirstAuthorGender),colour=conference))+ geom_boxplot(varwidth =
```

```
allData$stuGen = factor(paste(allData$conference,
                              allData$FirstAuthorGender),
                        levels=c("E8 F", "E8 M", "E9 F", "E9 M", "E10 F", "E10 M", "E11 F", "E11 M", "E12 F", "E12 M"))

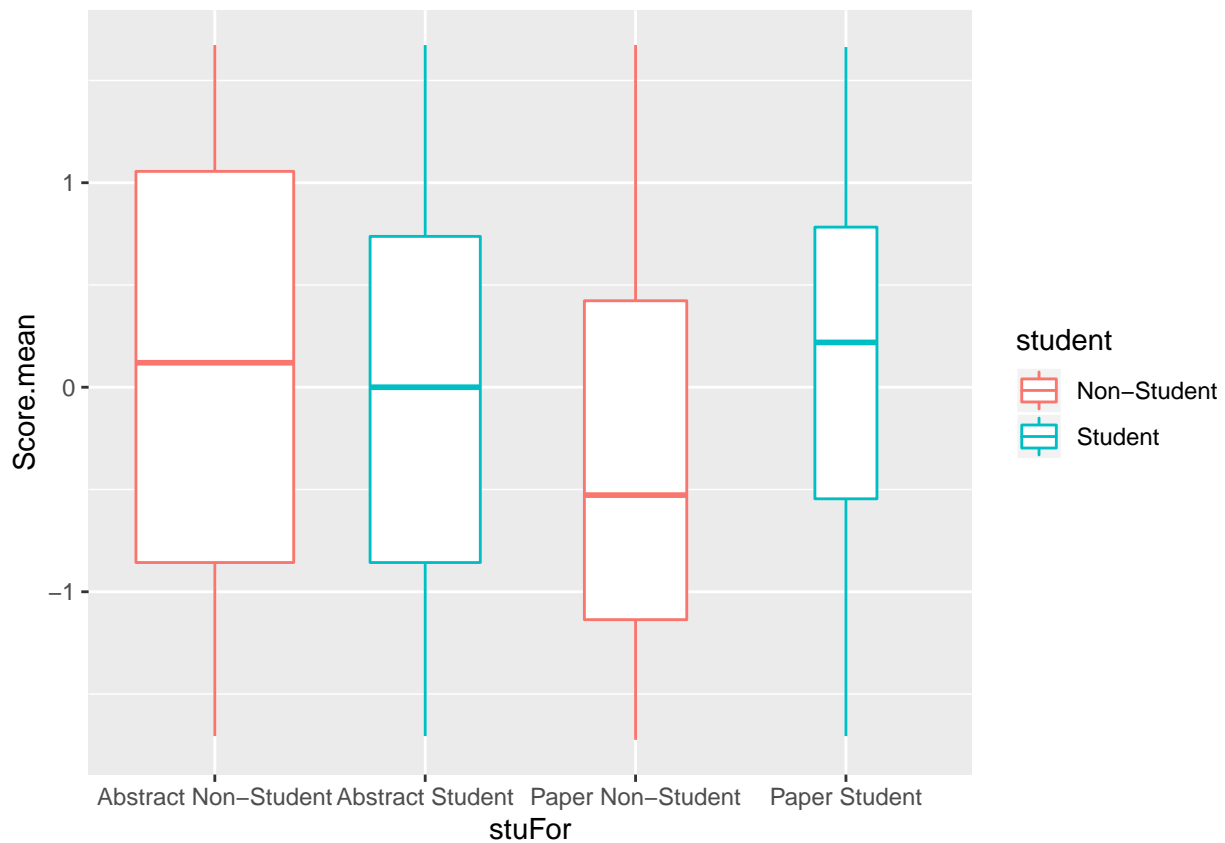
ad2 = allData[allData$conference!="E8",]

ggplot(ad2, mapping = aes(y=Score.mean,
                          x=stuGen,
                          colour=student))+
  geom_boxplot(varwidth = 0.5)
```



```
allData$stuFor = factor(paste(allData$format,
                              allData$student))

ggplot(allData[!is.na(allData$student),],
       mapping = aes(y=Score.mean,
                     x=stuFor,
                     colour=student))+
  geom_boxplot(varwidth = 0.5)
```



Summary statistics

```
t1 = table(allData$conference,allData$FirstAuthorGender)
t2 = table(allData$conference,allData$student)
t3 = table(allData$conference,allData$format)

cbind(t1,t2,t3)
```

```
##      F      M Non-Student Student Abstract Paper
## E8  58   94           0       0       98   55
## E9  52  130        119       63      121   61
## E10 67  124        149       42      131   60
## E11 76  119        118       77      145   50
## E12 84  122        147       59      161   45
```

```
stargazer(allData,type = 'text')
```

```
##
## =====
## Statistic   N    Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## -----
## Score.mean 927 -0.000  0.998   -1.724 -0.857   0.948   1.673
## year        927  3.128  1.393    1       2       4       5
## -----
```

Statistics for female authors:

```
stargazer(allData[allData$FirstAuthorGender=="F",],type = 'text')
```

```
##
```

```
## =====
## Statistic   N   Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## -----
## Score.mean 337 0.103  1.015   -1.706  -0.857   1.056   1.663
## year       337 3.226  1.421    1.000   2.000   4.000   5.000
## -----

# Statistics for male authors:
stargazer(allData[allData$FirstAuthorGender=="M",],type = 'text')

##
## =====
## Statistic   N   Mean  St. Dev.  Min   Pctl(25) Pctl(75)  Max
## -----
## Score.mean 589 -0.057  0.984   -1.724  -0.903   0.738   1.673
## year       589 3.076  1.374    1.000   2.000   4.000   5.000
## -----
```

Review ranks by gender and student status

Are papers with female first authors ranked higher than those with male first authors under double-blind review?

Using a simple anova, there's a significant interaction between gender and review type:

```
summary(aov(Score.mean ~ FirstAuthorGender*student*review*format,
            data=allData[allData$conference!="E8",]))
```

```
##                                Df Sum Sq Mean Sq F value
## FirstAuthorGender              1    5.4    5.366    5.551
## student                        1    0.4    0.423    0.438
## review                         1    0.1    0.054    0.056
## format                         1   11.7   11.747   12.151
## FirstAuthorGender:student       1    0.8    0.758    0.784
## FirstAuthorGender:review        1    4.3    4.278    4.425
## student:review                  1    0.3    0.302    0.313
## FirstAuthorGender:format        1    0.9    0.946    0.979
## student:format                  1   10.1   10.079   10.426
## review:format                   1    0.7    0.701    0.725
## FirstAuthorGender:student:review 1    0.0    0.005    0.005
## FirstAuthorGender:student:format 1    0.0    0.037    0.038
## FirstAuthorGender:review:format  1    0.3    0.270    0.279
## student:review:format            1    2.1    2.124    2.197
## FirstAuthorGender:student:review:format 1    0.1    0.080    0.082
## Residuals                      758  732.8    0.967
##                                Pr(>F)
## FirstAuthorGender          0.018726 *
## student                    0.508378
## review                      0.813575
## format                     0.000519 ***
## FirstAuthorGender:student  0.376058
## FirstAuthorGender:review    0.035743 *
## student:review              0.576264
## FirstAuthorGender:format    0.322788
## student:format              0.001296 **
## review:format               0.394665
## FirstAuthorGender:student:review 0.943998
## FirstAuthorGender:student:format 0.844520
## FirstAuthorGender:review:format  0.597387
## student:review:format          0.138730
## FirstAuthorGender:student:review:format 0.774242
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, it looks like this is driven just by EvoLang11:

```
t.test.string = function(tx){
  t = signif(tx$statistic,2)
  df = tx$parameter['df']
  p = signif(tx$p.value,3)
  est = signif(diff(tx$estimate),2)

  paste("(difference in means = ",est," t = ",t," p = ",p,")",sep = " ")
}
```

```

}
for(conf in levels(allData$conference)){
  print(conf)
  print(t.test.string(t.test(Score.mean~FirstAuthorGender, data=allData[allData$conference==conf,])))
}

```

```

## [1] "E8"
## [1] "(difference in means = -0.092, t = 0.54, p = 0.591)"
## [1] "E9"
## [1] "(difference in means = 0.14, t = -0.87, p = 0.386)"
## [1] "E10"
## [1] "(difference in means = -0.12, t = 0.75, p = 0.454)"
## [1] "E11"
## [1] "(difference in means = -0.61, t = 4.4, p = 1.93e-05)"
## [1] "E12"
## [1] "(difference in means = -0.058, t = 0.4, p = 0.687)"

```

There is also a significant main effect of first author gender.

The model above mots EvoLang 8 because it has no data for student status. We get the same results if we omit student status and run the test for all conferences:

```

summary(aov(Score.mean ~ FirstAuthorGender*review*format,
            data=allData))

```

```

##              Df Sum Sq Mean Sq F value    Pr(>F)
## FirstAuthorGender      1      5.5    5.480    5.603 0.01814 *
## review                  1      0.0    0.032    0.032 0.85706
## format                  1      8.6    8.649    8.843 0.00302 **
## FirstAuthorGender:review      1      4.9    4.852    4.961 0.02617 *
## FirstAuthorGender:format      1      2.4    2.439    2.494 0.11463
## review:format              1      1.6    1.641    1.678 0.19553
## FirstAuthorGender:review:format 1      0.0    0.019    0.019 0.89015
## Residuals                918   897.9    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness

```

Mixed effects model

Alternatively, we can use a mixed effects model, with random slopes for conference and test whether the interaction between gender and review type is a significant fixed predictor. A random intercept is not necessary, because the data is scaled to be centered around 0 within each conference. A random slope for the interaction between gender and review is also not permissible, since review type does not vary by conference.

```
contrasts(allData$FirstAuthorGender) <- contr.sum(2)/2
contrasts(allData$review) <- contr.sum(2)/2
contrasts(allData$student) <- contr.sum(2)/2
contrasts(allData$format) <- contr.sum(2)/2

m0 <- lmer(
  Score.mean ~
    1 + (FirstAuthorGender*review*student*format) +
    (0+FirstAuthorGender+student+format|conference),
  allData[allData$conference!="E8",],
  control=lmerControl(optimizer="bobyqa",optCtrl = list(maxfun=1000000)),
  REML = T
)

## boundary (singular) fit: see ?isSingular

summary(m0)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## Score.mean ~ 1 + (FirstAuthorGender * review * student * format) +
## (0 + FirstAuthorGender + student + format | conference)
## Data: allData[allData$conference != "E8", ]
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+07))
##
## REML criterion at convergence: 2175.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.0469 -0.8318 -0.0649  0.8731  2.1003
##
## Random effects:
##   Groups             Name                Variance Std.Dev. Corr
##   conference FirstAuthorGenderF 0.049878 0.22333
##              FirstAuthorGenderM 0.002765 0.05258 -0.97
##              student1           0.045642 0.21364 -0.87  0.73
##              format1            0.023844 0.15441  0.37 -0.14 -0.77
## Residual                    0.950489 0.97493
## Number of obs: 774, groups:  conference, 4
##
## Fixed effects:
##
##              Estimate Std. Error
## (Intercept) -0.005526  0.063844
## FirstAuthorGender1
## review1     -0.094290  0.127687
## student1    -0.203825  0.142736
## format1      0.154509  0.121783
```

```
## FirstAuthorGender1:review1      0.256651  0.332875
## FirstAuthorGender1:student1     -0.208867  0.189766
## review1:student1                0.217541  0.285473
## FirstAuthorGender1:format1      0.088045  0.188464
## review1:format1                 0.286881  0.243566
## student1:format1                0.620946  0.189427
## FirstAuthorGender1:review1:student1 0.070548  0.379532
## FirstAuthorGender1:review1:format1 0.178654  0.376927
## FirstAuthorGender1:student1:format1 0.250443  0.377860
## review1:student1:format1        -0.543252  0.378853
## FirstAuthorGender1:review1:student1:format1 0.151257  0.755720
##                                     df t value Pr(>|t|)
## (Intercept)                      2.851857 -0.087  0.9367
## FirstAuthorGender1                2.601224  0.882  0.4519
## review1                          2.851857 -0.738  0.5163
## student1                         2.859042 -1.428  0.2528
## format1                          3.396149  1.269  0.2845
## FirstAuthorGender1:review1        2.601224  0.771  0.5046
## FirstAuthorGender1:student1       674.800702 -1.101  0.2714
## review1:student1                 2.859042  0.762  0.5040
## FirstAuthorGender1:format1       749.272361  0.467  0.6405
## review1:format1                  3.396149  1.178  0.3148
## student1:format1                 615.328521  3.278  0.0011 **
## FirstAuthorGender1:review1:student1 674.800701  0.186  0.8526
## FirstAuthorGender1:review1:format1 749.272361  0.474  0.6357
## FirstAuthorGender1:student1:format1 719.938801  0.663  0.5077
## review1:student1:format1         615.328521 -1.434  0.1521
## FirstAuthorGender1:review1:student1:format1 719.938801  0.200  0.8414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

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

The results above suggest that there's no overall interaction between gender and review type. The tendency is there, but from the plots it's probably just driven by EvoLang 11.

We can run the same model without student status to include data from EvoLang 8:

```
m0 <- lmer(
  Score.mean ~
    1 + (FirstAuthorGender*review*format) +
    (0+FirstAuthorGender+format|conference),
  allData,
  control=lmerControl(optimizer="bobyqa",optCtrl = list(maxfun=1000000)),
  REML = T
)

## boundary (singular) fit: see ?isSingular
summary(m0)
```



```

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Score.mean ~ 1 + (FirstAuthorGender * review * format) + (0 +
##   FirstAuthorGender + format | conference)
##   Data: allData
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+07))
##
## REML criterion at convergence: 2616.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.00126 -0.87372 -0.03911  0.89455  2.01352
##
## Random effects:
##   Groups             Name                Variance Std.Dev. Corr
##   conference FirstAuthorGenderF 0.018667 0.13663
##              FirstAuthorGenderM 0.005532 0.07438  -0.61
##              format1            0.050963 0.22575  -0.39 -0.49
##   Residual                0.966469 0.98309
## Number of obs: 926, groups:  conference, 5
##
## Fixed effects:
##
##              Estimate Std. Error      df t value
## (Intercept)    -0.04969    0.04647   6.05500  -1.069
## FirstAuthorGender1    0.11629    0.11771   3.93339   0.988
## review1         -0.02834    0.09293   6.05500  -0.305
## format1          0.26421    0.12964   3.40816   2.038
## FirstAuthorGender1:review1    0.29146    0.23542   3.93339   1.238
## FirstAuthorGender1:format1    0.21076    0.15756  904.20405   1.338
## review1:format1    0.17404    0.25928   3.40816   0.671
## FirstAuthorGender1:review1:format1 -0.05484    0.31512  904.20405  -0.174
##
##              Pr(>|t|)
## (Intercept)      0.326
## FirstAuthorGender1    0.380
## review1           0.771
## format1           0.123
## FirstAuthorGender1:review1    0.284
## FirstAuthorGender1:format1    0.181
## review1:format1    0.545
## FirstAuthorGender1:review1:format1 0.862
##
## Correlation of Fixed Effects:
##              (Intr) FrsAG1 reviw1 formt1 FrstAthrGndr1:r1
## FrstAthrGn1      0.443
## review1          0.202  0.067
## format1         -0.606 -0.149 -0.165
## FrstAthrGndr1:r1 0.067  0.204  0.443 -0.033
## FrstAthrGndr1:f1 -0.197 -0.334 -0.044  0.206 -0.126
## revw1:frmt1     -0.165 -0.033 -0.606  0.202 -0.149
## FrstAG1:1:1     -0.044 -0.126 -0.197  0.019 -0.334
##
##              FrstAthrGndr1:f1 rvw1:1
## FrstAthrGn1
## review1
## format1

```

```
## FrstAthrGndr1:r1
## FrstAthrGndr1:f1
## revw1:frmt1      0.019
## FrstAG1:1:1      0.206      0.206
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

Again, there's no interaction between gender and review type.

Permutation test

The distributions of score means are not very normal within conferences. We run a permutation test to address this. We calculate the average difference between single blind and double blind scores for males (dM) and for females (dF). Then we calculate dF - dM. A value > 0 means females scores increase more than male scores under double blind review. This 'true difference' is compared to a 'permuted difference'. The association between review scores and review type is randomly permuted, and dF - dM is calculated again. This is done 10,000 times to compare the true difference to a distribution of random differences.

```
meanDifferenceBetweenGenders = function(d){
  # difference in means between review types
  # for males
  # (change from single to double)
  diffMales = diff(rev(tapply(d[d$FirstAuthorGender=="M"],)$Score.mean,
    d[d$FirstAuthorGender=="M"],$review,
    mean)))
  # for females
  diffFemales = diff(rev(tapply(d[d$FirstAuthorGender=="F"],)$Score.mean,
    d[d$FirstAuthorGender=="F"],$review,
    mean)))
  # difference in differences
  # value > 0 means female scores increase
  # more under double-blind review than male scores
  return(diffFemales-diffMales)
}

perm = function(d){
  d$review = sample(d$review)
  meanDifferenceBetweenGenders(d)
}

perm.test = function(d,title){
  n = 10000
  trueDiff = meanDifferenceBetweenGenders(d)
  permDiff = replicate(n, perm(d))

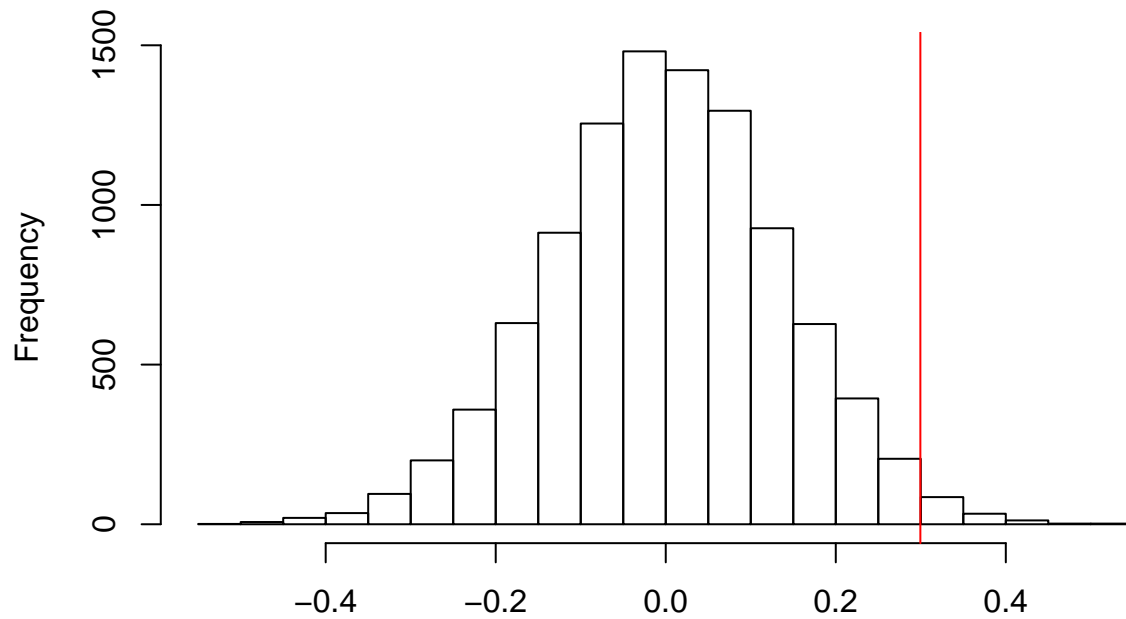
  p = sum(permDiff>trueDiff) / n
  z = (trueDiff-mean(permDiff)) / sd(permDiff)
  print(paste("p=",p," , z=",z))
  hist(permDiff,xlab="Female advantage in double-blind",main=title)
  abline(v=trueDiff,col=2)
}
```

Permutation test for all data:

```
perm.test(allData,
  "All conferences")
```

```
## [1] "p= 0.0135 , z= 2.18104312416607"
```

All conferences



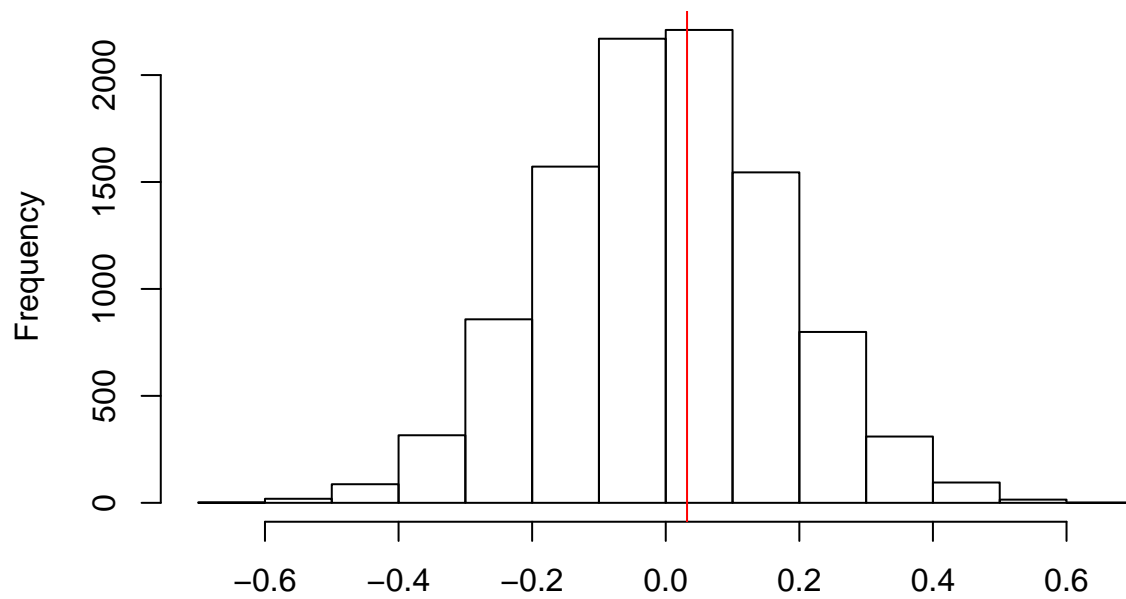
Female advantage in double-blind

Permutation test without E11 data:

```
perm.test(allData[allData$conference!="E11",],  
          "Without E11")
```

```
## [1] "p= 0.4216 , z= 0.196161240316842"
```

Without E11

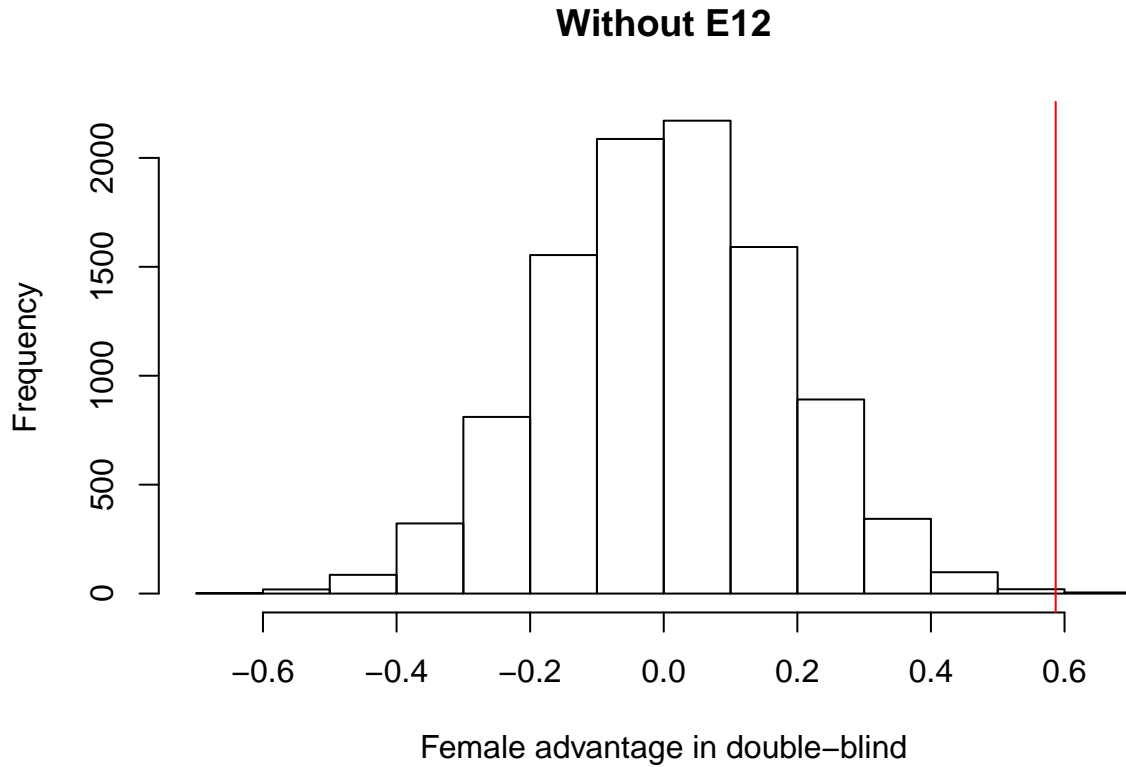


Female advantage in double-blind

Permutation test without E12 data:

```
perm.test(allData[allData$conference!="E12",],  
          "Without E12")
```

```
## [1] "p= 7e-04 , z= 3.30130459572563"
```

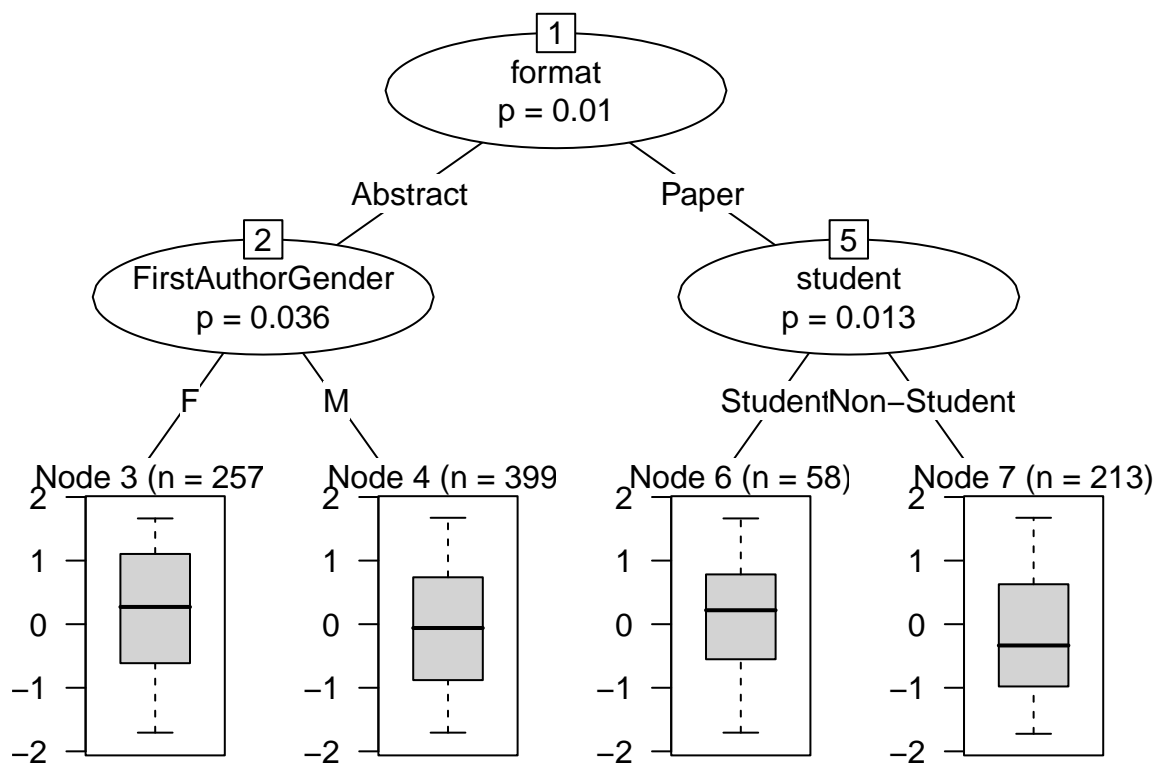


The results are in line with the test above. Across the whole data, females are given higher scores in double-blind, but this is driven by E11 alone.

Decision tree exploration

Construct a decision tree, attempting to predict review scores by format, student status, gender, review model and conference.

```
set.seed(2389)
for(f in c("conference", "format", "student", "FirstAuthorGender", "review")){
  allData[,f] = as.factor(allData[,f])
}
ct = ctree(Score.mean ~ format + student +
  FirstAuthorGender + review + conference, data=allData)
plot(ct)
```



Work out differences between leaves of the tree:

```
paperVabstract = tapply(allData$Score.mean, allData$format, mean)
paperVabstract
```

```
##      Abstract      Paper
## 0.06519752 -0.15782129
```

```
pStudentVpNonStudent = tapply(allData[
  allData$format=="Paper",]$Score.mean,
  allData[allData$format=="Paper",]$student, mean)
pStudentVpNonStudent
```

```
## Non-Student      Student
## -0.3300312  0.1235369
```

The tree suggests that full papers are given lower ratings than abstracts on average (about 6.6% difference). For full papers, students are given higher ratings than non-students (about 13.4% difference).

Readability scores

This section uses the file `EvoLang_ReadingScores_E8_to_E12.csv`. It includes the following variables:

- `conference`: Conference
- `gender`: Gender of first author
- `student`: Student status
- `format`: Full paper or short abstract
- `char_count`, `word_count`, `sent_count`, `sybl_count`: Number of characters, words, sentences and syllables. These distributions have been scaled and centred.
- `*_score`: Various measures of readability, calculated using the tools from Hengel (2016).
- `Score.mean`: Mean raw score given by reviewers (scaled between 0 and 1, higher = better paper)

Read the data:

```
readScores = read.csv("../data/EvoLang_ReadingScores_E8_to_E12.csv", stringsAsFactors = F)
```

We'll focus on the Flesch-Kincaid score (since most other measures are highly correlated with it and it's easy to interpret) and the Dale-Chall score (which is not highly correlated with the other measures):

```
round(cor(readScores[,c("flesch_score", "fleschkincaid_score",  
                        "gunningfog_score", "smog_score", "dalechall_score")]), 2)
```

```
##                flesch_score fleschkincaid_score gunningfog_score  
## flesch_score                1.00                -0.91                -0.90  
## fleschkincaid_score          -0.91                1.00                0.98  
## gunningfog_score             -0.90                0.98                1.00  
## smog_score                   -0.93                0.96                0.99  
## dalechall_score              -0.73                0.62                0.62  
##                smog_score dalechall_score  
## flesch_score              -0.93              -0.73  
## fleschkincaid_score        0.96                0.62  
## gunningfog_score           0.99                0.62  
## smog_score                 1.00                0.65  
## dalechall_score            0.65                1.00
```

Scale the variables:

```
readScores$fleschkincaid_score_scaled = scale(readScores$fleschkincaid_score)  
readScores$dalechall_score_scaled = scale(readScores$dalechall_score)  
readScores$student[readScores$student=="EC"] = "Non-Student"  
readScores$student[readScores$student=="Faculty"] = "Non-Student"  
# Remove an outlier  
readScores = readScores[readScores$fleschkincaid_score_scaled < 6,]  
readScores$gender = factor(readScores$gender)  
  
readScores$conference = factor(readScores$conference,  
                               levels = c("E8", "E9", "E10", "E11", "E12"))  
  
# Box-Cox scaling  
pp = preprocess(readScores[,  
                 c('fleschkincaid_score', 'dalechall_score')],  
                method="BoxCox")  
lambda.fk = pp$bc$fleschkincaid_score$lambda  
lambda.dc = pp$bc$dalechall_score$lambda  
readScores$fleschkincaid_score_norm =
```

```

bcPower(readScores$fleschkincaid_score, lambda = lambda.fk)
readScores$dalechall_score_norm =
  bcPower(readScores$dalechall_score, lambda = lambda.dc)
readScores$Score.mean.norm = scale(readScores$Score.mean)

readScores$review = factor(c("Single", "Double")[(readScores$conference %in% c("E11", "E12"))+1])
readScores$student = factor(readScores$student)
readScores$format = factor(readScores$format)

```

Create time variable: a continuous variable increasing with each conference.

```
readScores$time = as.numeric(readScores$conference)-3
```

Number of available datapoints (less than the total because some papers could not be automatically converted to text):

```
table(readScores$conference, readScores$gender)
```

```
##
##           F    M
##  E8      56   93
##  E9      52  130
##  E10     67  120
##  E11     68  111
##  E12     84  121
```

```

gtable2 = table(readScores$gender, readScores$conference, readScores$student)
write.csv(cbind(t(gtable2[, , 1]), t(gtable2[, , 2])),
          "../results/CountTable_Readability.csv")
gtable2

```

```

## , , = Non-Student
##
##
##      E8 E9 E10 E11 E12
##  F   0 34  55  38  54
##  M   0 85  90  72  92
##
## , , = Student
##
##
##      E8 E9 E10 E11 E12
##  F   0 18  12  30  30
##  M   0 45  30  39  29

```

Flesch-Kinkaid score

Descriptive stats

```
mean(readScores$fleschkincaid_score)
```

```
## [1] 13.16953
```

```
cor.test(readScores$fleschkincaid_score, readScores$dalechall_score)
```

```

##
## Pearson's product-moment correlation

```



```
##
## data: readScores$fleschkincaid_score and readScores$dalechall_score
## t = 24.187, df = 900, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.5864068 0.6656599
## sample estimates:
## cor
## 0.6276569

sel = readScores$conference=="E11"
mean(readScores[sel & readScores$gender=="M",]$fleschkincaid_score) -
  mean(readScores[sel & readScores$gender=="F",]$fleschkincaid_score)

## [1] -0.4849936

sel = readScores$conference=="E12"
mean(readScores[sel & readScores$gender=="M",]$fleschkincaid_score) -
  mean(readScores[sel & readScores$gender=="F",]$fleschkincaid_score)

## [1] 0.1066118

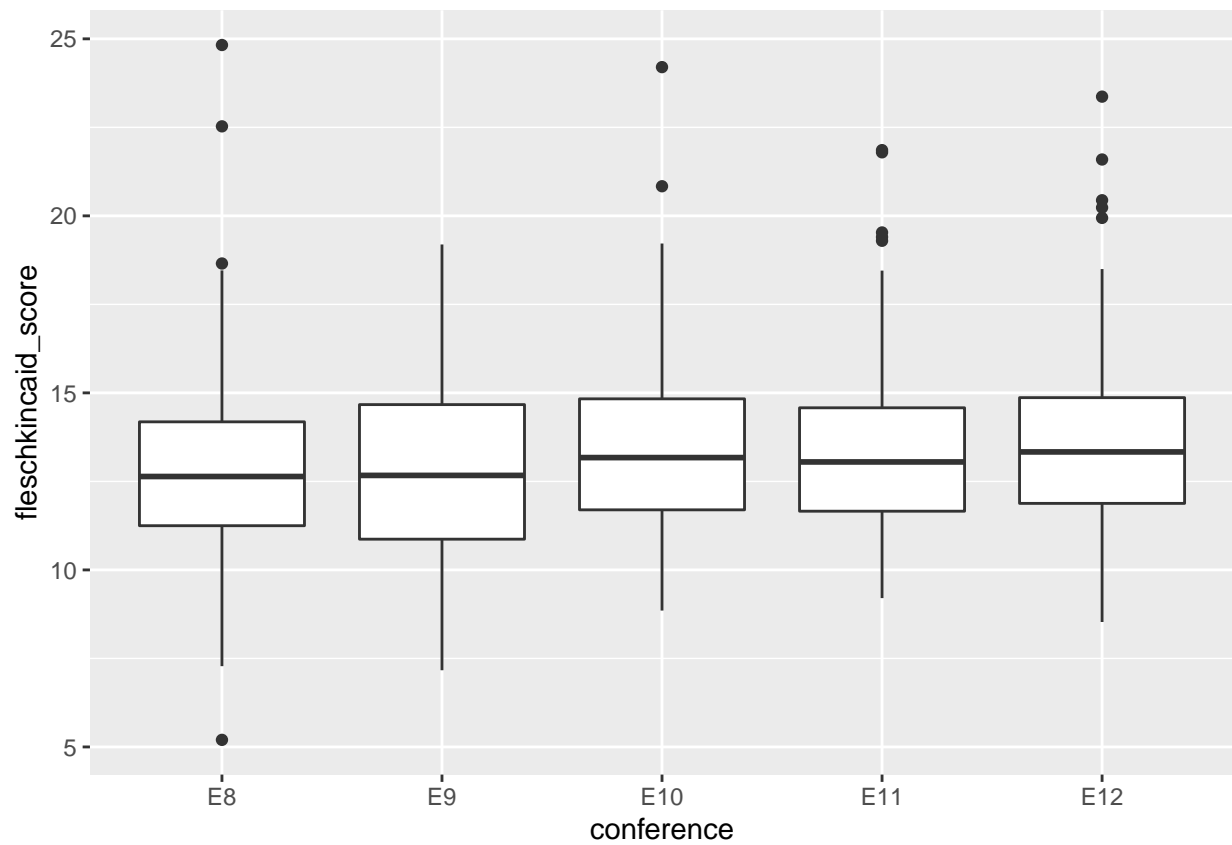
meanFK =
  rbind(tapply(readScores$fleschkincaid_score[readScores$gender=="F"],
    readScores$conference[readScores$gender=="F"],mean),
  tapply(readScores$fleschkincaid_score[readScores$gender=="M"],
    readScores$conference[readScores$gender=="M"],mean))
sdFK =
  rbind(tapply(readScores$fleschkincaid_score[readScores$gender=="F"],
    readScores$conference[readScores$gender=="F"],sd),
  tapply(readScores$fleschkincaid_score[readScores$gender=="M"],
    readScores$conference[readScores$gender=="M"],sd))

msdFK = matrix(paste0(round(meanFK,2), " (",round(sdFK,2),")"),nrow=2)
colnames(msdFK) = sort(unique(readScores$conference))
rownames(msdFK) = c("Female","Male")
write.csv(msdFK,"../results/MeanFleschKincaidScores_by_conf_by_gender.csv")
```

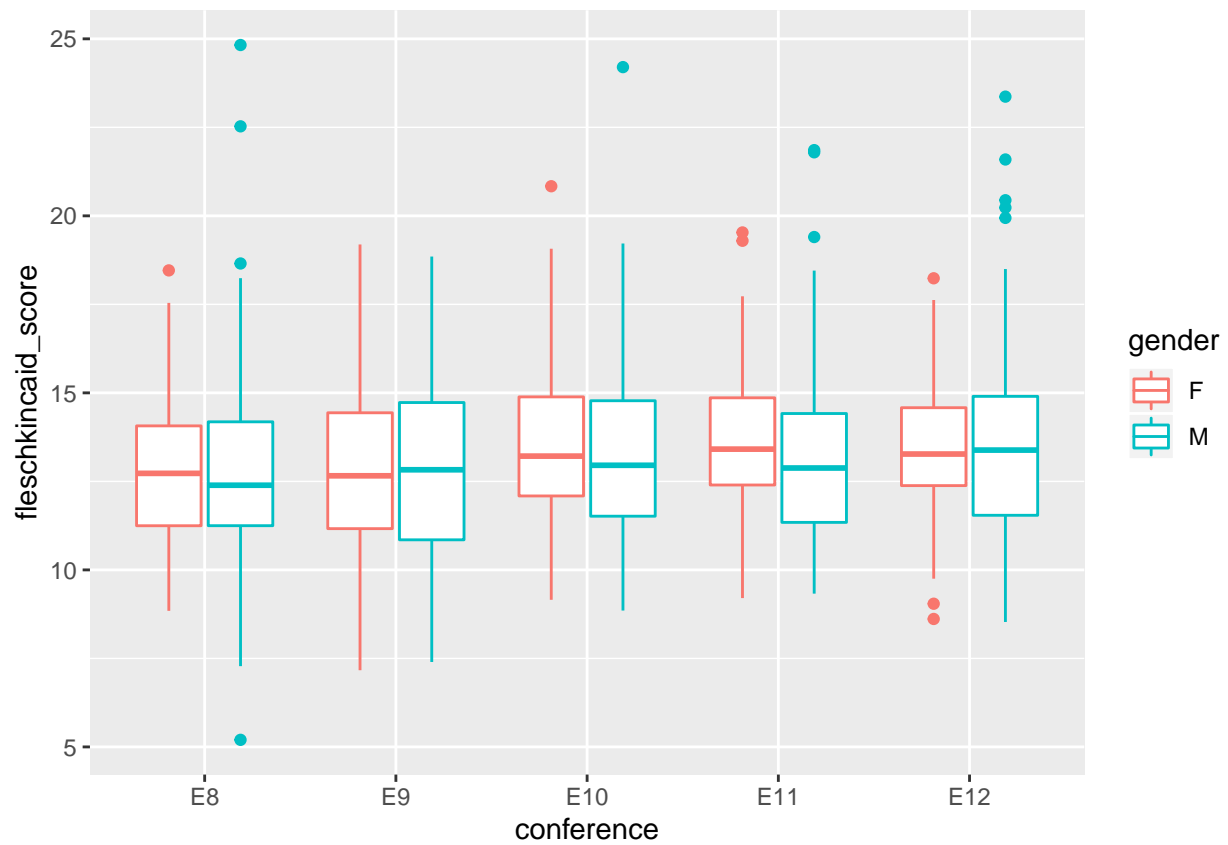
Various Plots:

```
readScores$gender2 = "Female"
readScores$gender2[readScores$gender=="M"] = "Male"

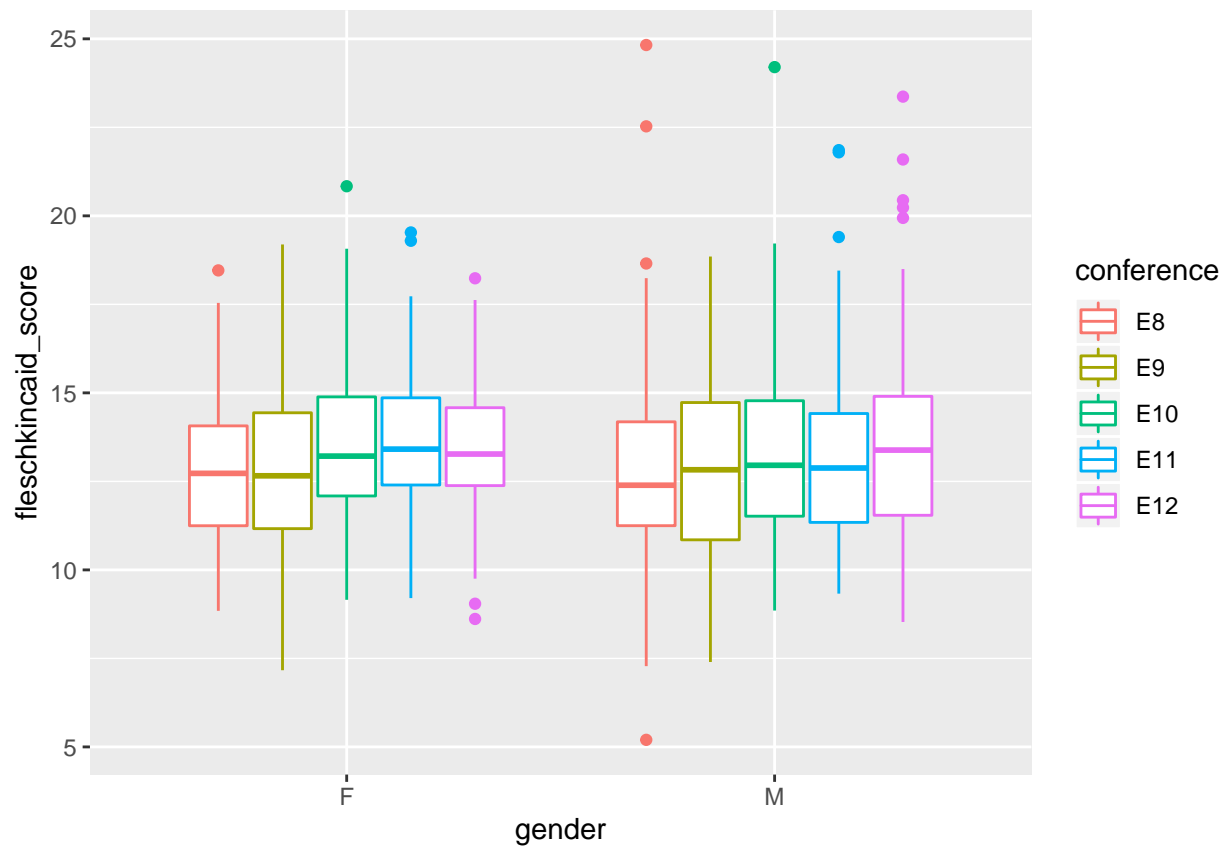
ggplot(readScores, aes(y=fleschkincaid_score,x=conference)) + geom_boxplot()
```



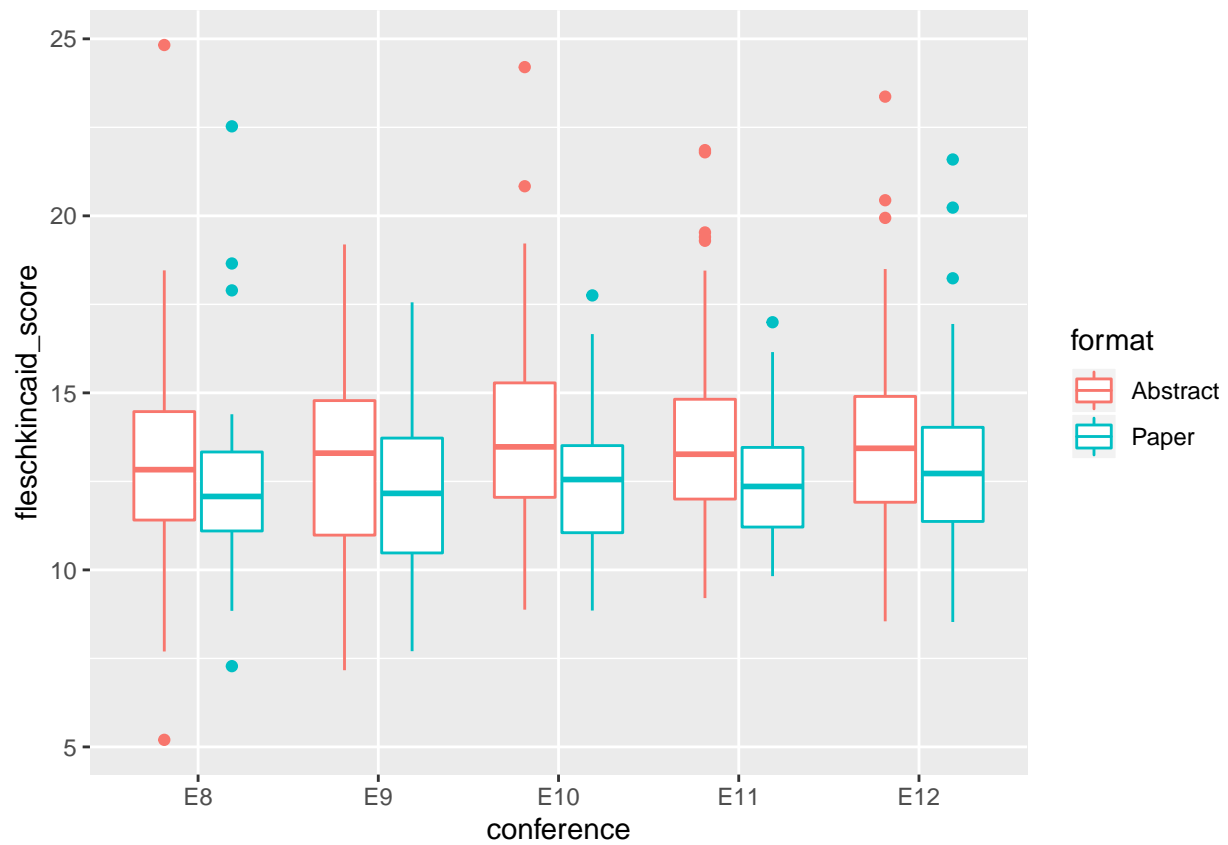
```
ggplot(readScores, aes(y=fleschkincaid_score,x=conference,colour=gender)) + geom_boxplot()
```



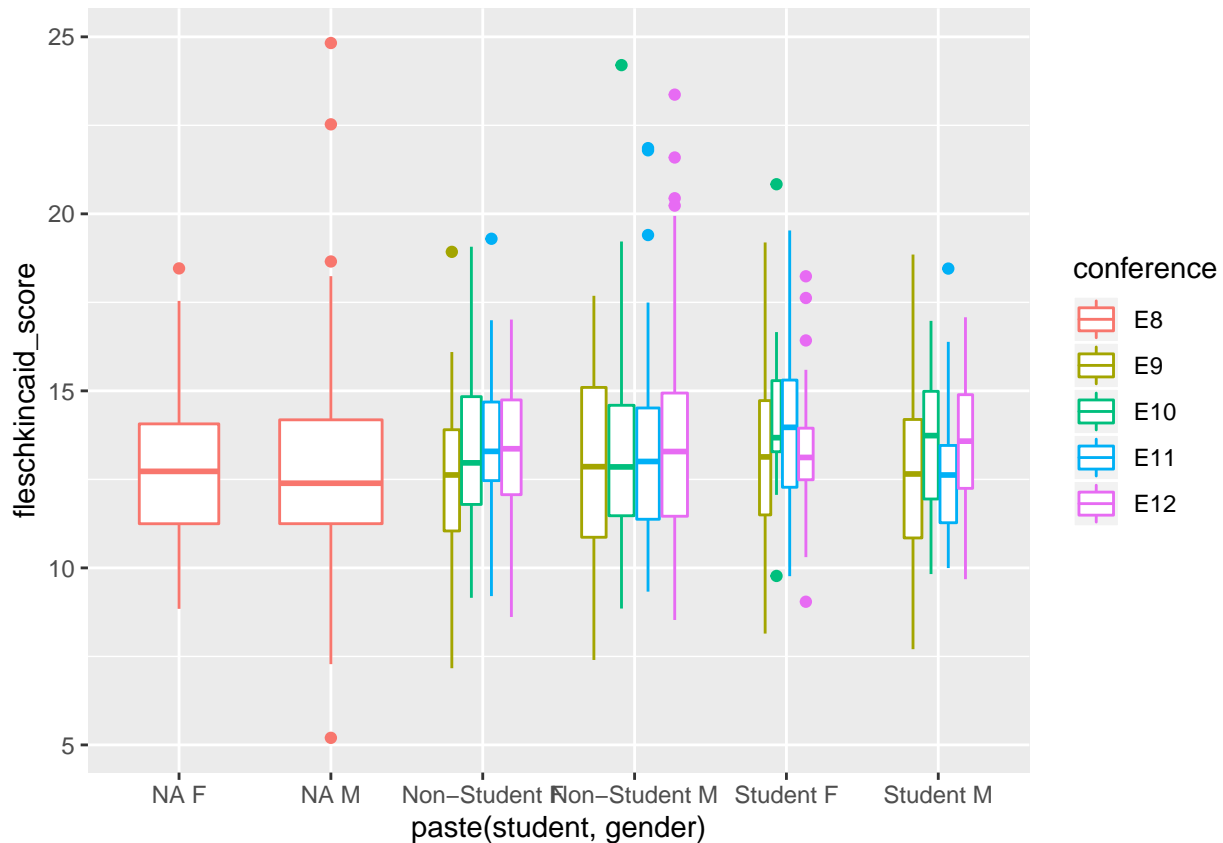
```
ggplot(readScores, aes(y=fleschkincaid_score,x=gender,colour=conference)) + geom_boxplot()
```



```
ggplot(readScores, aes(y=fleschkincaid_score,x=conference,colour=format)) + geom_boxplot()
```



```
ggplot(readScores, aes(y=fleschkincaid_score,x=paste(student,gender),colour=conference))+ geom_boxplot(
```



```
fkrs = ggplot(readScores, aes(y=fleschkincaid_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Flesch-Kincaid reading score", x="Gender")

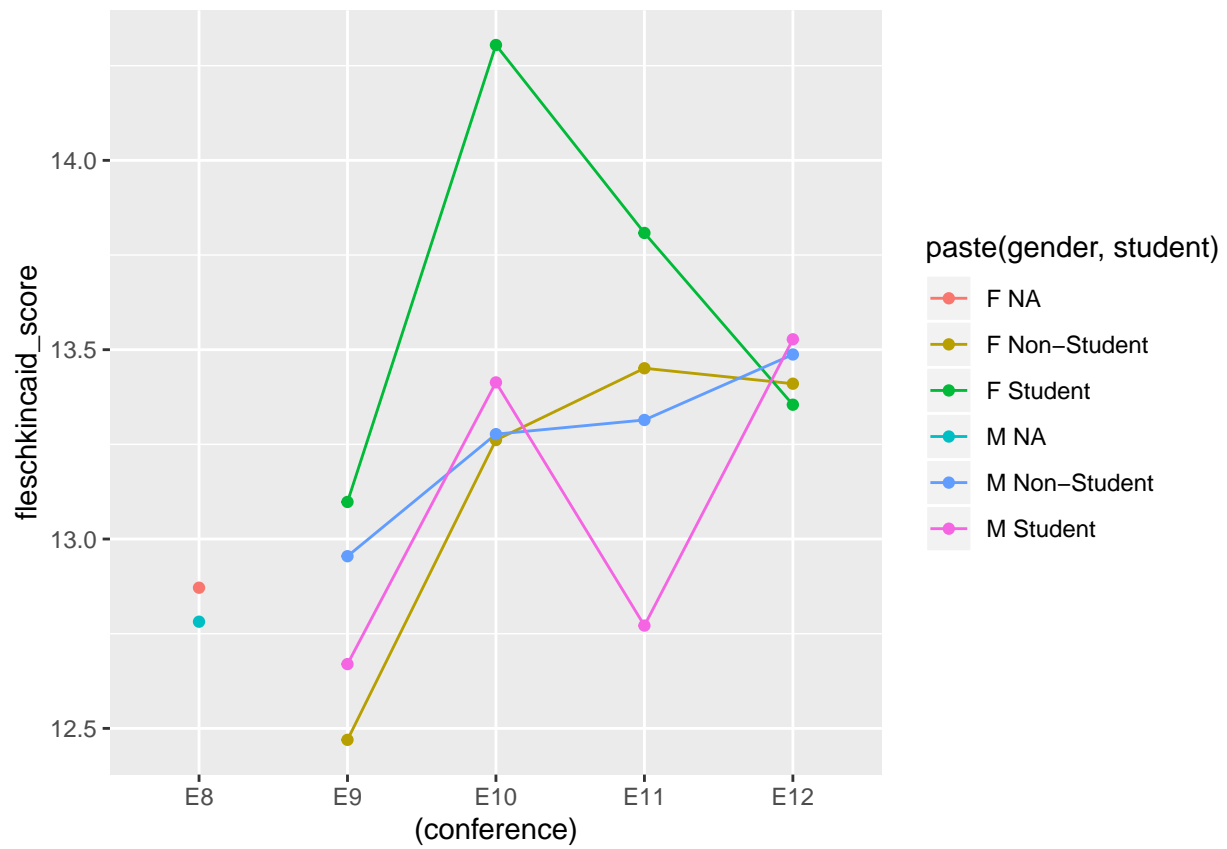
pdf("../results/FleschKincaidReadingScores.pdf",
     width=6,height=4)
fkrs
dev.off()
```

```
## pdf
## 2
```

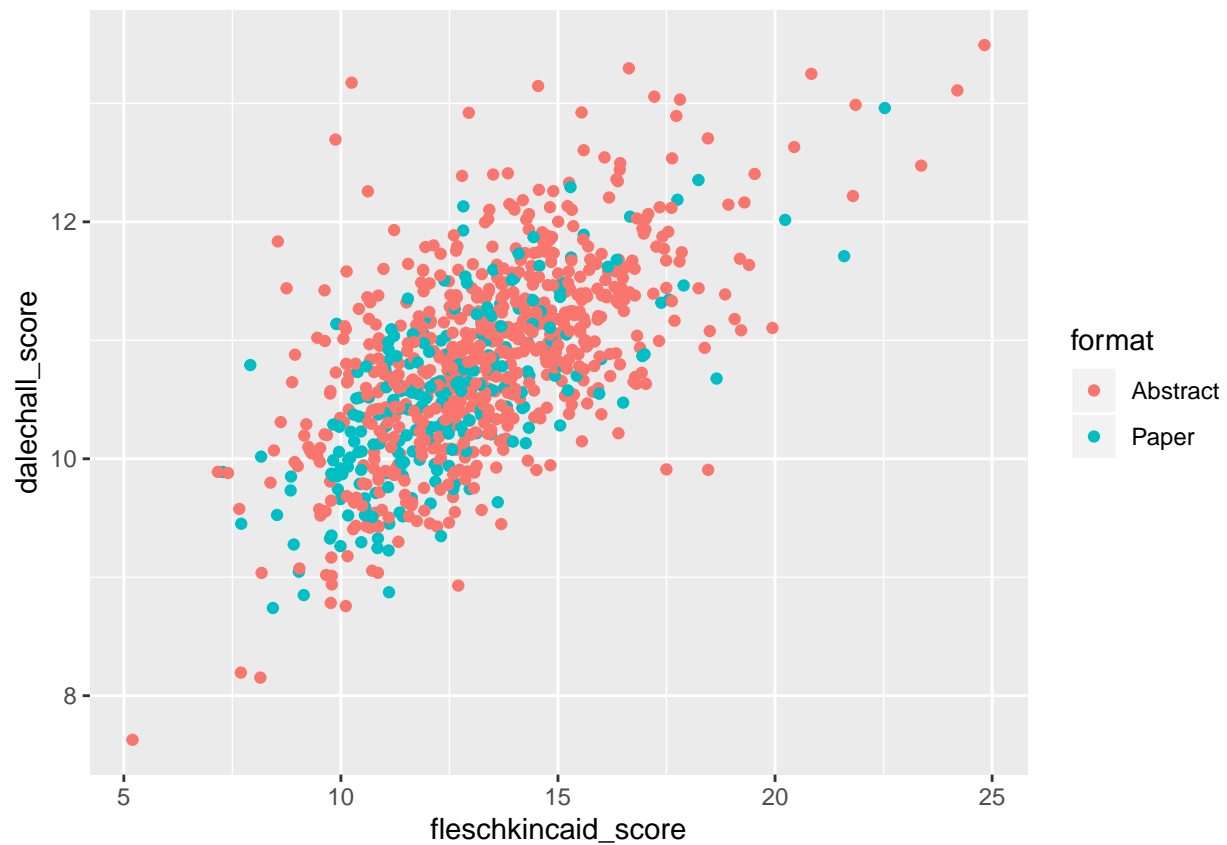
```
x = readScores %>% group_by(conference,gender,student) %>%
  summarise(dalechall_score=mean(dalechall_score),
            fleschkincaid_score=mean(fleschkincaid_score))
```

```
## Warning: Factor `student` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
ggplot(x,aes(x=(conference),y=fleschkincaid_score,
            group=paste(gender,student),
            colour=paste(gender,student))) +
  geom_line() + geom_point()
```



```
ggplot(readScores,
  aes(x=fleschkincaid_score,
    y=dalechall_score,
    colour=format)) +
  geom_point()
```



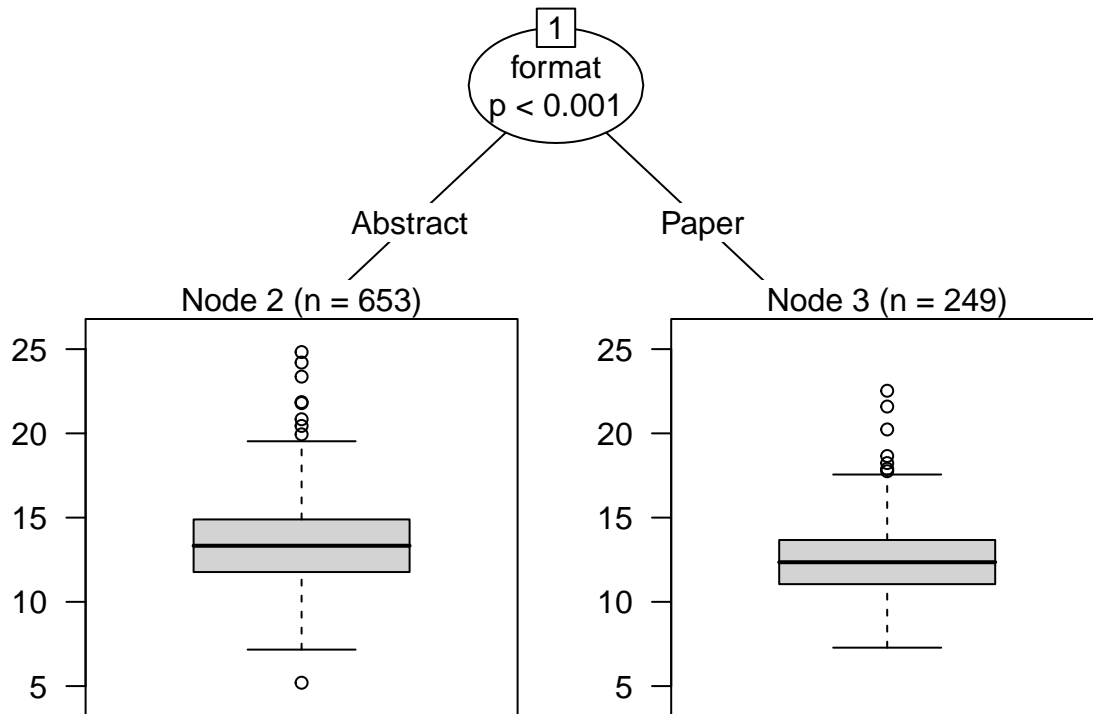
```
fkr2= ggplot(readScores, aes(y=fleschkincaid_score,
                             x=conference,colour=gender)) +
  geom_boxplot() +
  labs(y="Flesch-Kincaid reading score", x="Conference")

pdf("../results/FleschKincaidReadingScores2.pdf",
     width=6,height=4)
fkr2
dev.off()
```

```
## pdf
## 2
```

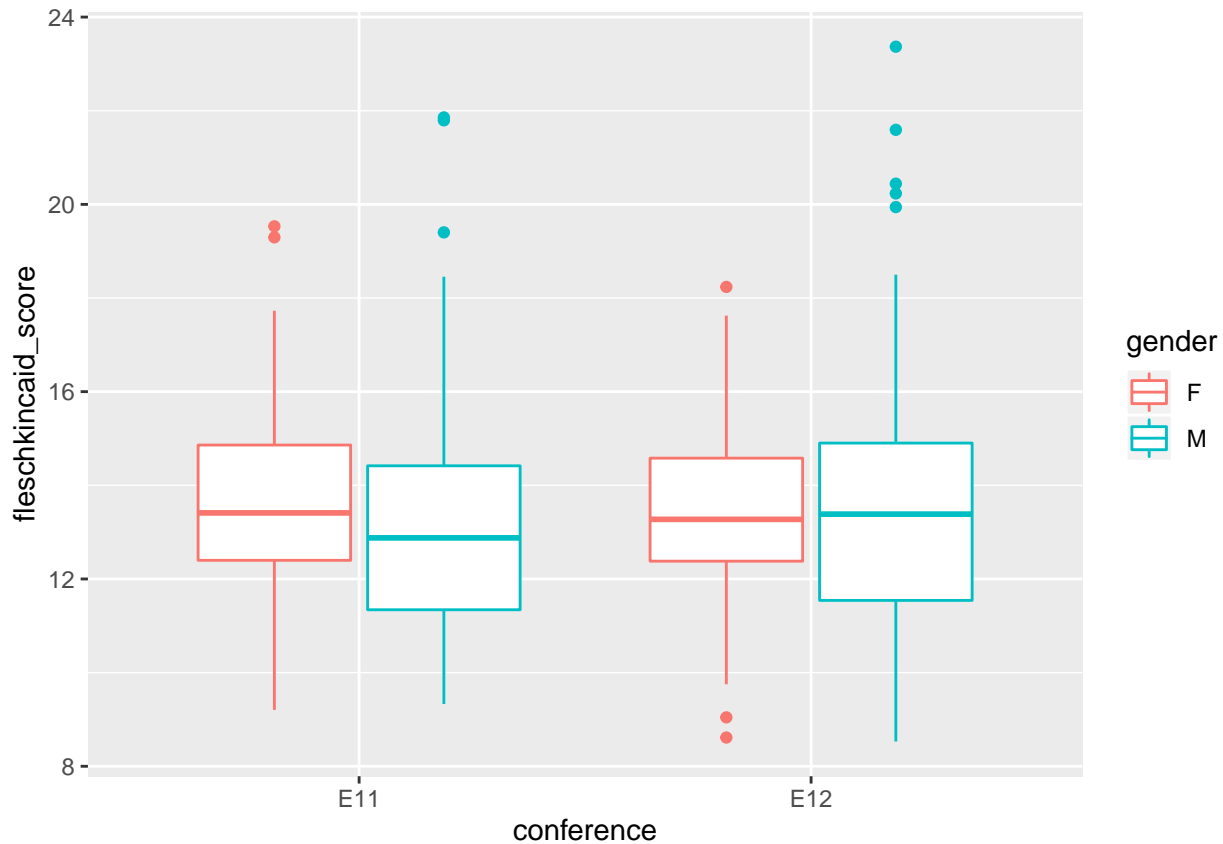

Decision tree

```
plot(ctree(fleschkincaid_score~  
  review+gender+time+format,  
  data=readScores))
```



Is there a gender difference between E11 and E12?

```
ggplot(readScores[readScores$conference %in% c("E11","E12"),],
       aes(x = conference, y=fleschkincaid_score, colour=gender)) +
  geom_boxplot()
```



```
summary(aov(fleschkincaid_score_norm~
  format*conference*student*gender,
  data = readScores[readScores$conference %in% c("E11","E12"),]))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------------------------------|-----|--------|---------|---------|------------|
| ## format | 1 | 0.812 | 0.8124 | 9.955 | 0.00174 ** |
| ## conference | 1 | 0.012 | 0.0120 | 0.147 | 0.70126 |
| ## student | 1 | 0.027 | 0.0271 | 0.332 | 0.56499 |
| ## gender | 1 | 0.026 | 0.0262 | 0.321 | 0.57110 |
| ## format:conference | 1 | 0.079 | 0.0794 | 0.973 | 0.32455 |
| ## format:student | 1 | 0.026 | 0.0264 | 0.323 | 0.57017 |
| ## conference:student | 1 | 0.032 | 0.0317 | 0.389 | 0.53344 |
| ## format:gender | 1 | 0.164 | 0.1642 | 2.012 | 0.15691 |
| ## conference:gender | 1 | 0.048 | 0.0485 | 0.594 | 0.44145 |
| ## student:gender | 1 | 0.063 | 0.0631 | 0.774 | 0.37970 |
| ## format:conference:student | 1 | 0.070 | 0.0702 | 0.860 | 0.35442 |
| ## format:conference:gender | 1 | 0.000 | 0.0000 | 0.000 | 0.98360 |
| ## format:student:gender | 1 | 0.001 | 0.0006 | 0.007 | 0.93366 |
| ## conference:student:gender | 1 | 0.077 | 0.0775 | 0.949 | 0.33058 |
| ## format:conference:student:gender | 1 | 0.210 | 0.2102 | 2.576 | 0.10937 |
| ## Residuals | 368 | 30.031 | 0.0816 | | |
| ## --- | | | | | |

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is an effect for format, but nothing else.

Mixed effects model across the whole readability data. The model was not converging with a random slope for student, so:

```
contrasts(readScores$gender) <- contr.sum(2)/2
contrasts(readScores$student) <- contr.sum(2)/2
contrasts(readScores$format) <- contr.sum(2)/2

m0 = lmer(fleschkincaid_score_scaled~ 1 +
          (format+student+gender+review)^2 + time +
          (1 + format + student + gender | conference),
          data = readScores[readScores$conference!="E8",])
```

boundary (singular) fit: see ?isSingular

```
summary(m0)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: fleschkincaid_score_scaled ~ 1 + (format + student + gender +
## review)^2 + time + (1 + format + student + gender | conference)
## Data: readScores[readScores$conference != "E8", ]
##
## REML criterion at convergence: 2048.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.4852 -0.6903 -0.0664  0.6061  4.4238
##
## Random effects:
##   Groups      Name      Variance Std.Dev. Corr
## conference (Intercept) 4.141e-03 0.064348
##           format1      5.417e-03 0.073598  1.00
##           student1     4.563e-05 0.006755 -1.00 -1.00
##           gender1      5.717e-03 0.075612  1.00  1.00 -1.00
## Residual              8.625e-01 0.928724
## Number of obs: 753, groups: conference, 4
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   -0.20220    0.15330   3.12138  -1.319   0.2756
## format1        0.29502    0.13732   8.71872   2.148   0.0612 .
## student1       0.06377    0.11725  343.30024   0.544   0.5869
## gender1        0.15884    0.12648   6.82810   1.256   0.2504
## reviewSingle   0.18933    0.20748   3.05860   0.913   0.4276
## time           0.12880    0.08792   2.93717   1.465   0.2410
## format1:student1 -0.02854    0.17551  736.11676  -0.163   0.8709
## format1:gender1  -0.28256    0.16945  736.84210  -1.668   0.0958 .
## format1:reviewSingle 0.02376    0.17156   5.32824   0.139   0.8949
## student1:gender1  -0.27867    0.15495  727.98045  -1.798   0.0725 .
## student1:reviewSingle -0.14903    0.15106  245.41848  -0.987   0.3248
## gender1:reviewSingle -0.06184    0.16340   4.74401  -0.378   0.7214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
```

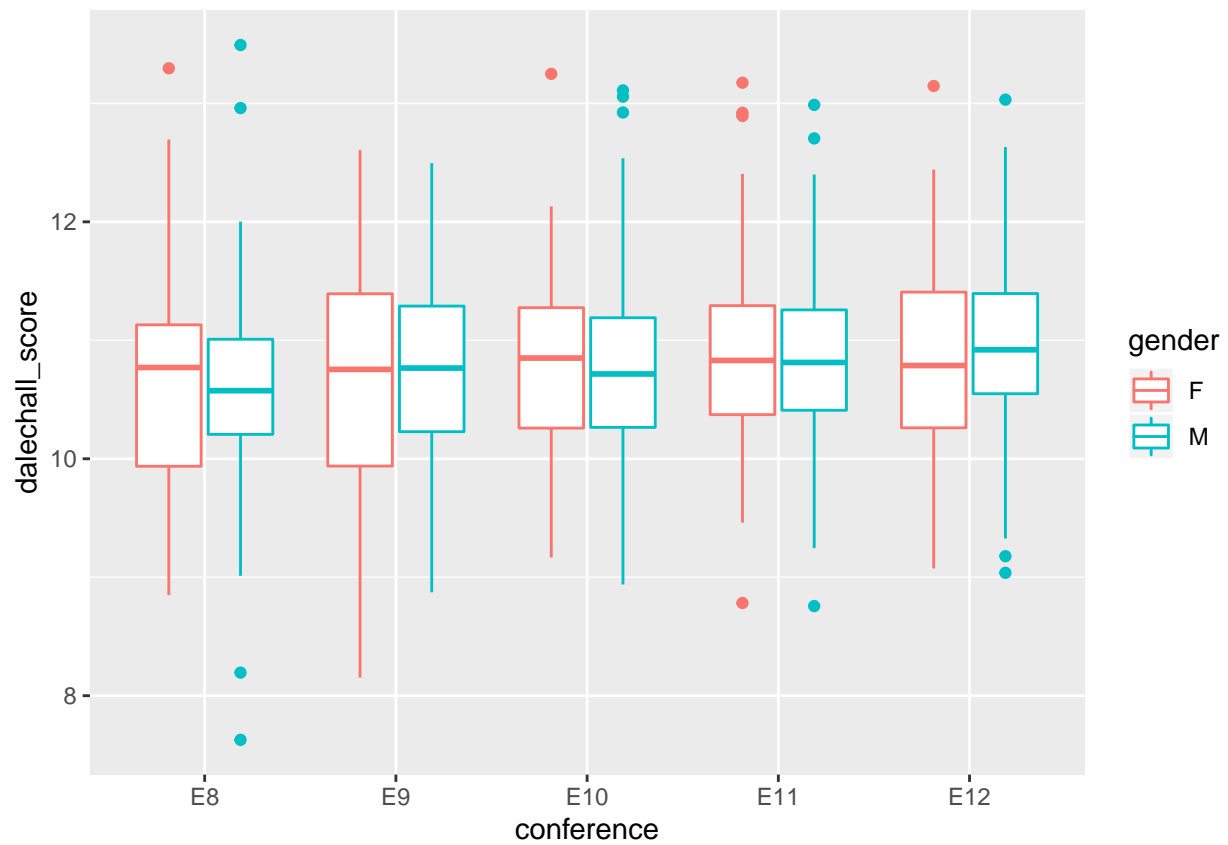
```
##          (Intr) format1 stdnt1 gendr1 rvwSng time   frmt1:s1 frmt1:g1
## format1      -0.104
## student1     -0.113  0.210
## gender1       0.261 -0.031  0.032
## reviewSingl  -0.918  0.044  0.043 -0.173
## time         -0.854 -0.033 -0.068 -0.036  0.851
## frmt1:stdn1   0.073 -0.307 -0.481 -0.003  0.002  0.041
## frmt1:gndr1  -0.147  0.240 -0.018 -0.428  0.076  0.061  0.030
## frmt1:rvwSn   0.049 -0.679 -0.059 -0.059 -0.028  0.014  0.015  0.003
## stdnt1:gnd1  -0.013  0.015  0.126 -0.203 -0.032  0.023 -0.078  0.059
## stdnt1:rvwS   0.119 -0.090 -0.634 -0.058 -0.182 -0.010  0.121  0.001
## gndr1:rvwSn  -0.152 -0.026 -0.040 -0.637  0.245 -0.010 -0.002  0.101
##          frm1:S std1:1 std1:S
## format1
## student1
## gender1
## reviewSingl
## time
## frmt1:stdn1
## frmt1:gndr1
## frmt1:rvwSn
## stdnt1:gnd1  0.006
## stdnt1:rvwS  0.055  0.125
## gndr1:rvwSn  0.113 -0.074 -0.017
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

Abstracts have higher reading scores than papers (marginally), but there are no other significant effects.

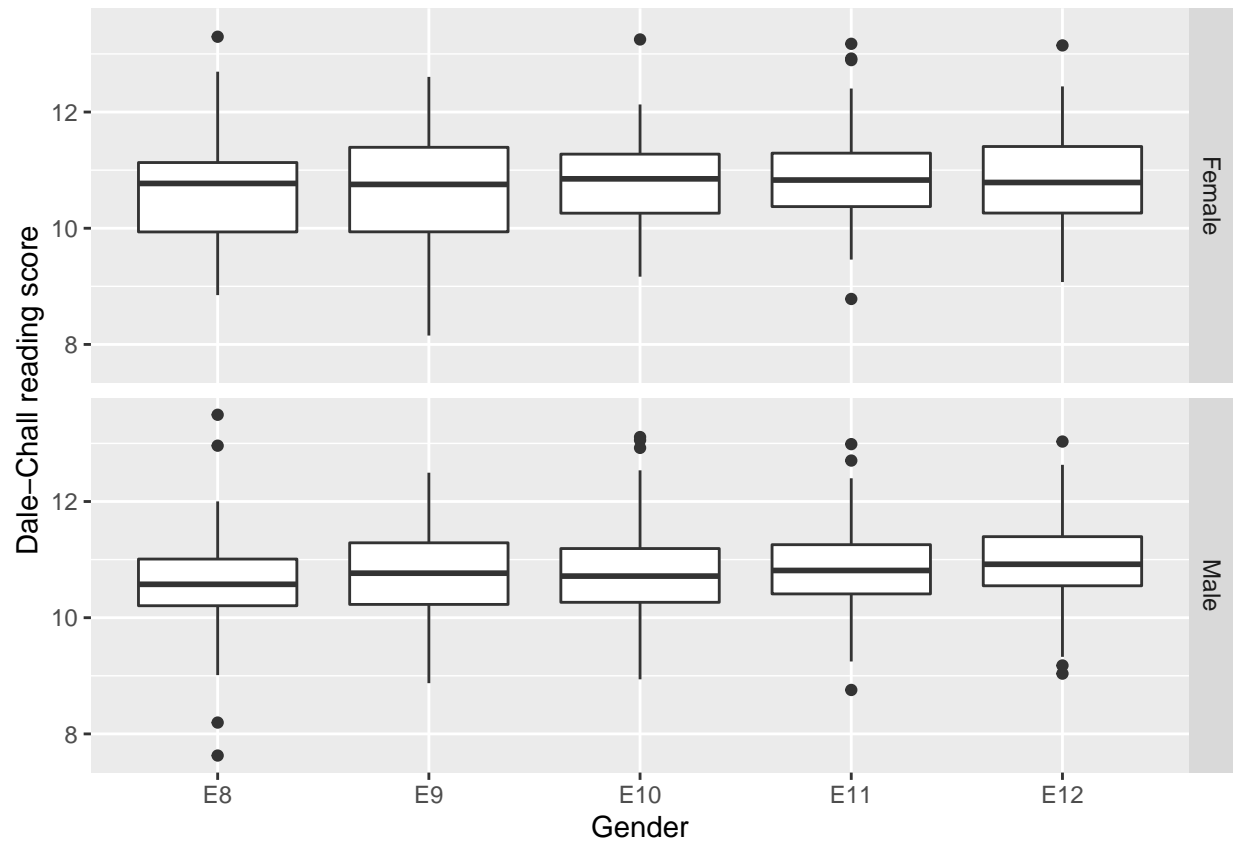
Dale-Chall scale

Plots

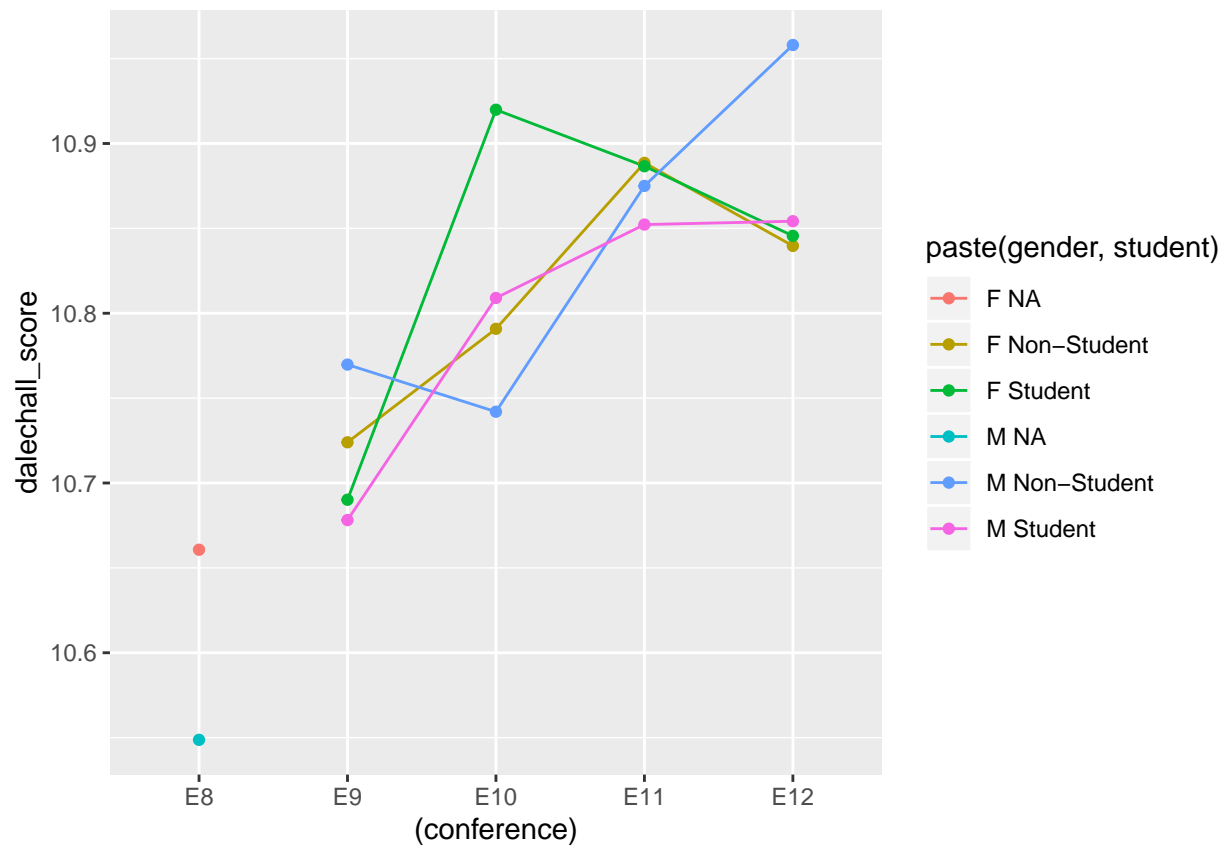
```
ggplot(readScores, aes(y=dalechall_score,x=conference,colour=gender)) + geom_boxplot()
```



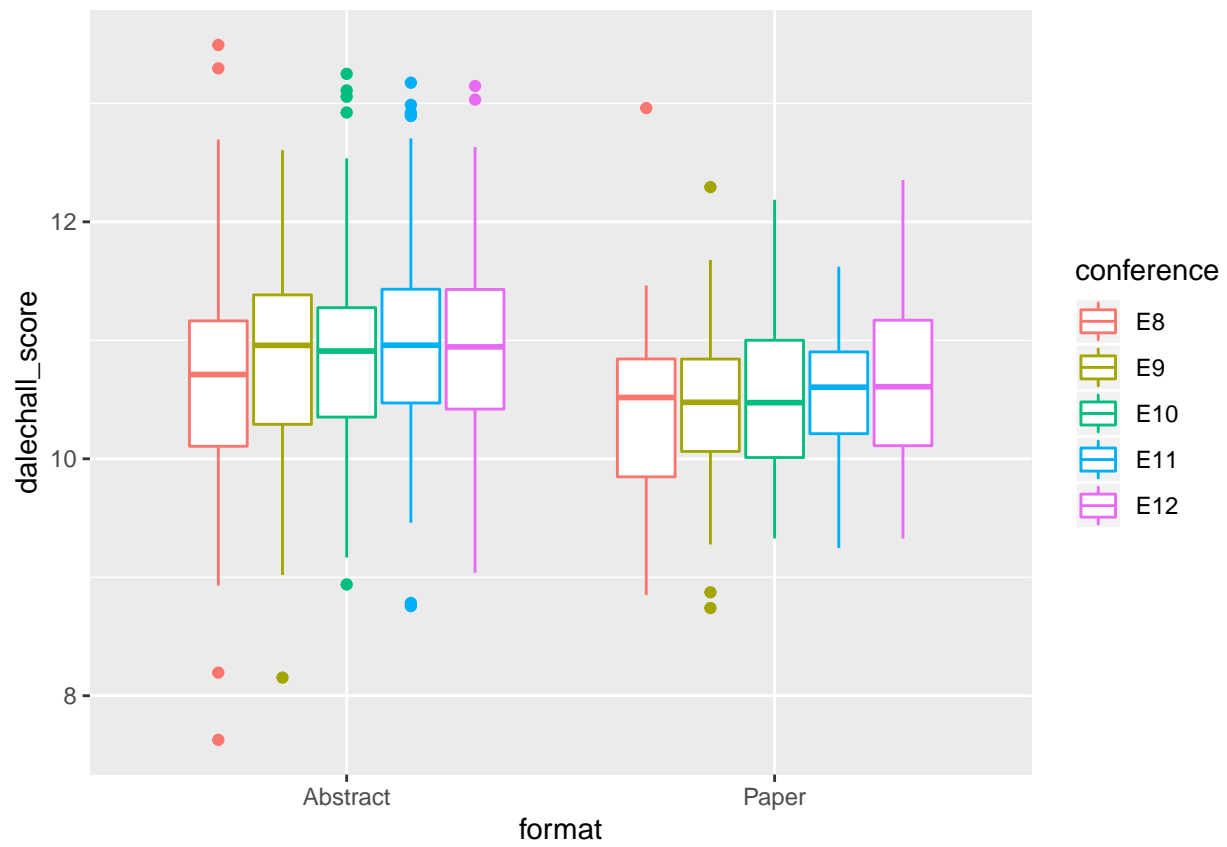
```
ggplot(readScores, aes(y=dalechall_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Dale-Chall reading score", x="Gender")
```



```
ggplot(x,aes(x=(conference),y=dalechall_score,group=paste(gender,student),colour=paste(gender,student)))
```

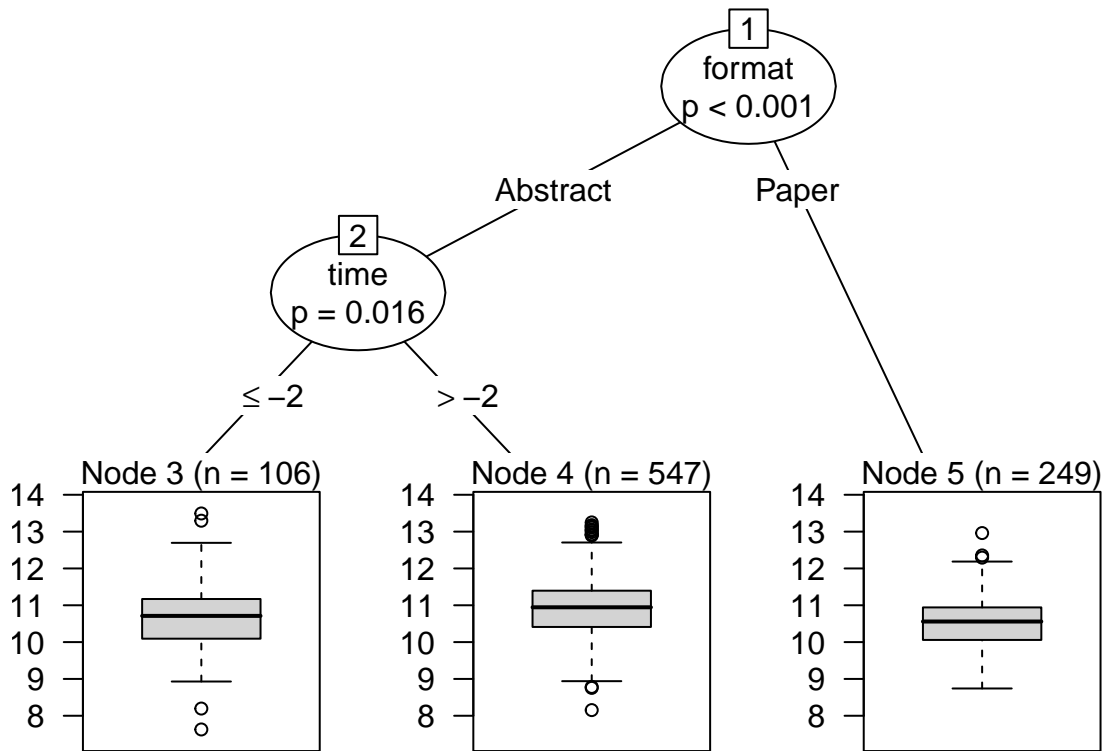


```
ggplot(readScores, aes(y=dalechall_score,x=format,colour=conference)) + geom_boxplot()
```

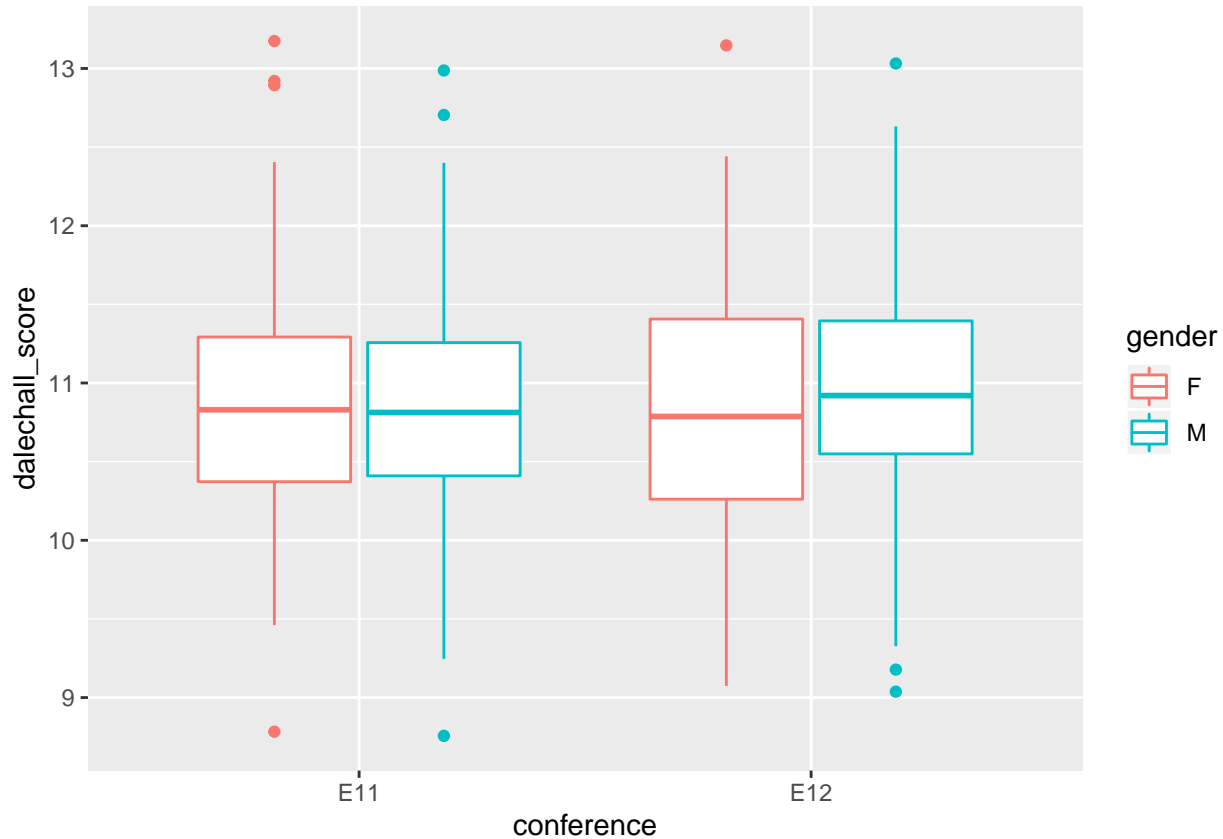
Decision tree:

```
plot(ctree(dalechall_score~review+gender+  
time+format,data=readScores))
```



Is there a gender difference between E11 and E12?

```
ggplot(readScores[readScores$conference %in% c("E11","E12"),],
       aes(x = conference, y=dalechall_score, colour=gender)) +
  geom_boxplot()
```



```
summary(aov(dalechall_score_norm~
  format*conference*student*gender,
  data = readScores[readScores$conference %in% c("E11","E12"),]))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-------------------------------------|-----|--------|---------|---------|--------------|
| ## format | 1 | 2.12 | 2.1248 | 14.947 | 0.000131 *** |
| ## conference | 1 | 0.00 | 0.0007 | 0.005 | 0.944331 |
| ## student | 1 | 0.14 | 0.1358 | 0.956 | 0.328937 |
| ## gender | 1 | 0.15 | 0.1488 | 1.047 | 0.306879 |
| ## format:conference | 1 | 0.08 | 0.0805 | 0.566 | 0.452138 |
| ## format:student | 1 | 0.19 | 0.1903 | 1.339 | 0.247965 |
| ## conference:student | 1 | 0.00 | 0.0000 | 0.000 | 0.985759 |
| ## format:gender | 1 | 0.07 | 0.0709 | 0.499 | 0.480522 |
| ## conference:gender | 1 | 0.02 | 0.0156 | 0.110 | 0.740284 |
| ## student:gender | 1 | 0.07 | 0.0744 | 0.523 | 0.469874 |
| ## format:conference:student | 1 | 0.13 | 0.1345 | 0.946 | 0.331379 |
| ## format:conference:gender | 1 | 0.01 | 0.0079 | 0.056 | 0.813858 |
| ## format:student:gender | 1 | 0.04 | 0.0378 | 0.266 | 0.606347 |
| ## conference:student:gender | 1 | 0.02 | 0.0204 | 0.143 | 0.705069 |
| ## format:conference:student:gender | 1 | 0.02 | 0.0184 | 0.130 | 0.718911 |
| ## Residuals | 368 | 52.31 | 0.1422 | | |
| ## --- | | | | | |

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There's an effect for format, but nothing else.

Mixed effects model across whole data:

Run mixed effects model:

```
m0 = lmer(dalechall_score_norm~ 1 +
          (format+student+gender+review)^2 + time +
          (1 + format + gender | conference),
          data = readScores[readScores$conference!="E8",])

## boundary (singular) fit: see ?isSingular

summary(m0)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## dalechall_score_norm ~ 1 + (format + student + gender + review)^2 +
##   time + (1 + format + gender | conference)
##   Data: readScores[readScores$conference != "E8", ]
##
## REML criterion at convergence: 709.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.6468 -0.6509  0.0085  0.6174  2.9993
##
## Random effects:
##   Groups      Name      Variance Std.Dev.  Corr
##   conference (Intercept) 0.000e+00 0.000e+00
##             format1     3.728e-12 1.931e-06   NaN
##             gender1     1.523e-10 1.234e-05   NaN 0.16
##   Residual              1.418e-01 3.766e-01
## Number of obs: 753, groups:  conference, 4
##
## Fixed effects:
##              Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)    6.093860   0.049515 740.999792 123.072 < 2e-16
## format1        0.178149   0.051526 740.999287   3.457 0.000576
## student1       0.016439   0.047485 741.000000   0.346 0.729303
## gender1       -0.037135   0.046453 740.981266  -0.799 0.424308
## reviewSingle  -0.018070   0.066442 740.999836  -0.272 0.785720
## time          0.009118   0.027792 740.999677   0.328 0.742950
## format1:student1 0.050858   0.071084 740.999999   0.715 0.474544
## format1:gender1 0.006176   0.068640 741.000000   0.090 0.928328
## format1:reviewSingle -0.016687 0.062840 740.998706  -0.266 0.790666
## student1:gender1 -0.042414 0.062770 740.999992  -0.676 0.499441
## student1:reviewSingle -0.021888 0.061028 741.000000  -0.359 0.719955
## gender1:reviewSingle 0.040176 0.058667 740.970473   0.685 0.493672
##
## (Intercept)      ***
## format1          ***
## student1
## gender1
## reviewSingle
## time
## format1:student1
```

```

## format1:gender1
## format1:reviewSingle
## student1:gender1
## student1:reviewSingle
## gender1:reviewSingle
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) formt1 stdnt1 gendr1 rvwSng time   frmt1:s1 frmt1:g1
## format1      -0.306
## student1     -0.134  0.244
## gender1       0.177 -0.228  0.054
## reviewSingl  -0.913  0.187  0.049 -0.103
## time         -0.844 -0.032 -0.080 -0.040  0.845
## frmt1:stdn1   0.120 -0.331 -0.482 -0.004 -0.026  0.018
## frmt1:gndr1  -0.155  0.260 -0.017 -0.472  0.067  0.043  0.029
## frmt1:rvwSn   0.198 -0.674 -0.078  0.085 -0.231  0.015  0.017  0.003
## stdnt1:gnd1   0.013  0.016  0.127 -0.223 -0.069 -0.005 -0.080  0.057
## stdnt1:rvwS   0.137 -0.111 -0.634 -0.076 -0.213 -0.011  0.123 -0.001
## gndr1:rvwSn  -0.078  0.120 -0.059 -0.620  0.146 -0.013 -0.001  0.114
##          frm1:S std1:1 std1:S
## format1
## student1
## gender1
## reviewSingl
## time
## frmt1:stdn1
## frmt1:gndr1
## frmt1:rvwSn
## stdnt1:gnd1  0.007
## stdnt1:rvwS  0.081  0.125
## gndr1:rvwSn -0.108 -0.084  0.001
## convergence code: 0
## boundary (singular) fit: see ?isSingular

```

Differences by format, but no other effects.

Reading scores and review scores

The simple correlations between reading score and review scores are weak, but suggest that higher scores are given to submissions with higher reading grades:

```
cor.test(readScores$Score.mean, readScores$fleschkincaid_score)
```

```

##
## Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$fleschkincaid_score
## t = 2.493, df = 900, p-value = 0.01285
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01763462 0.14729407
## sample estimates:

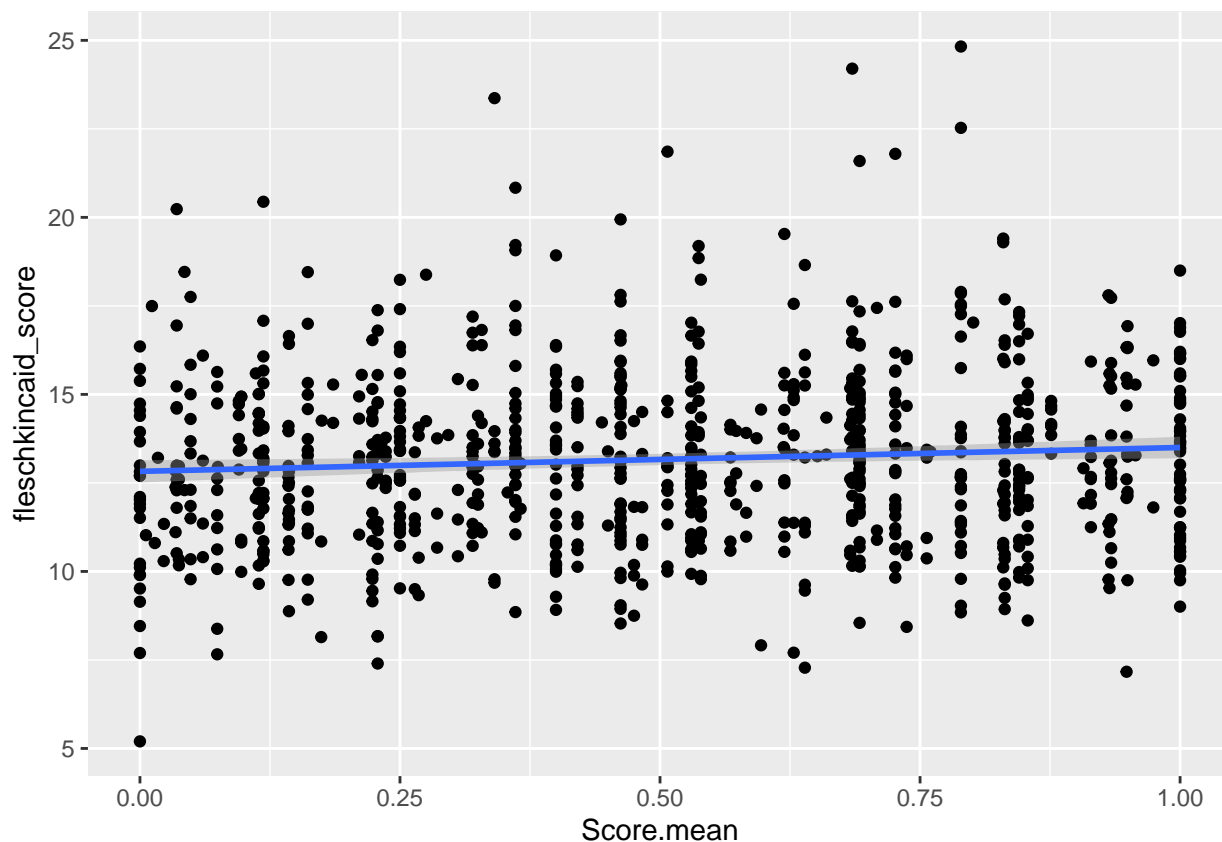
```

```
##      cor
## 0.0828148

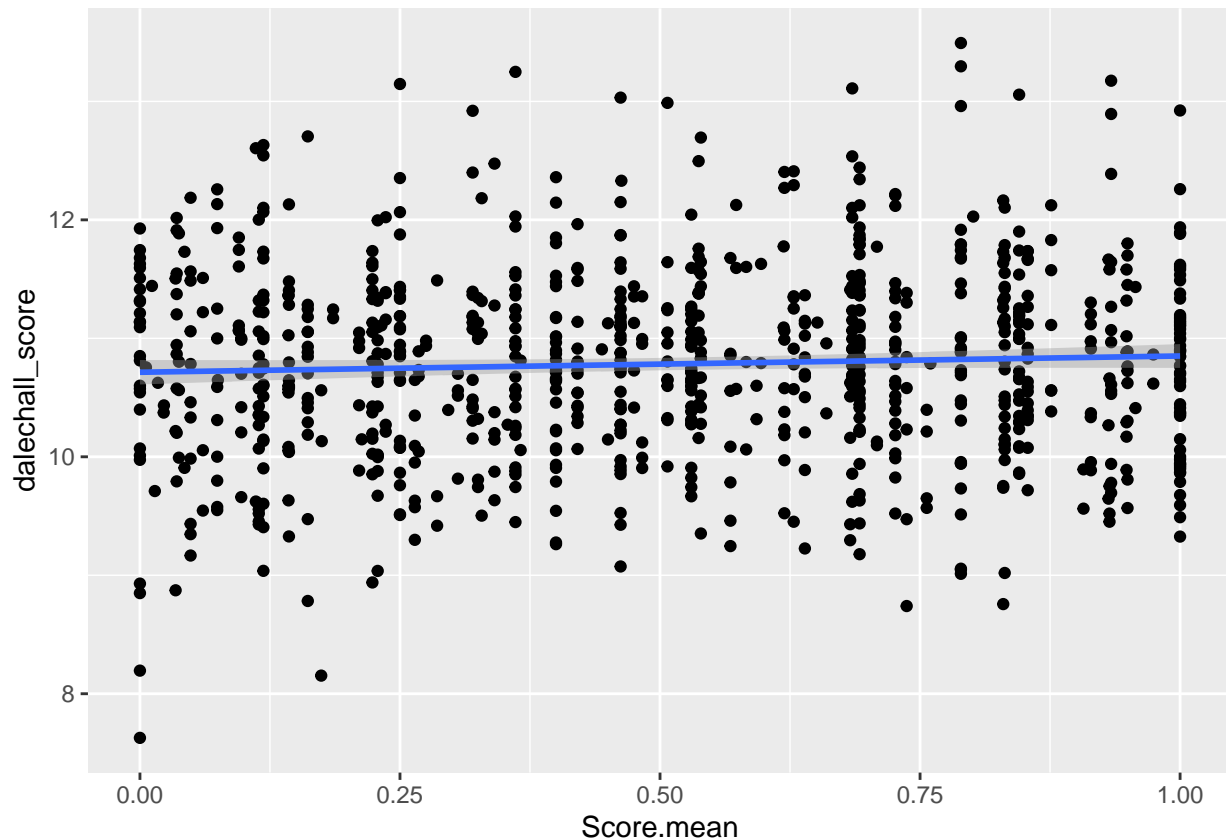
cor.test(readScores$Score.mean, readScores$dalechall_score)

##
## Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$dalechall_score
## t = 1.544, df = 900, p-value = 0.1229
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.0139239  0.1162837
## sample estimates:
##      cor
## 0.05139832
```

```
ggplot(readScores,
       aes(y=fleschkincaid_score,
           x=Score.mean)) +
  geom_point() +
  stat_smooth(method = 'lm')
```



```
ggplot(readScores,
       aes(y=dalechall_score,
           x=Score.mean)) +
  geom_point() +
  stat_smooth(method = 'lm')
```



Are there interactions between reading scores and gender?

```
m0 = lmer(Score.mean.norm ~ 1 +
          format + student + gender +
          (1 | conference),
          data = readScores,
          control = lmerControl(optimizer = 'Nelder_Mead'),
          REML = F)
```

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

```
m1 = update(m0, ~. + fleschkincaid_score_scaled)
```

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

```
m2 = update(m1, ~. + fleschkincaid_score_scaled:gender)
```

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

```
anova(m0, m1, m2)
```

```
## Data: readScores
```

```
## Models:
```

```
## m0: Score.mean.norm ~ 1 + format + student + gender + (1 | conference)
```

```
## m1: Score.mean.norm ~ format + student + gender + (1 | conference) +
```

```
## m1:   fleschkincaid_score_scaled
```

```
## m2: Score.mean.norm ~ format + student + gender + (1 | conference) +
```

```
## m2:   fleschkincaid_score_scaled + gender:fleschkincaid_score_scaled
```

```
##   Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
```

```
## m0  6 2126.2 2153.9 -1057.1   2114.2
```



```
## m1  7 2126.7 2159.1 -1056.4  2112.7 1.4054      1    0.2358
## m2  8 2128.7 2165.7 -1056.4  2112.7 0.0079      1    0.9292
```

```
summary(m2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
## Satterthwaite's method [lmerModLmerTest]
## Formula: Score.mean.norm ~ format + student + gender + (1 | conference) +
## fleschkincaid_score_scaled + gender:fleschkincaid_score_scaled
## Data: readScores
## Control: lmerControl(optimizer = "Nelder_Mead")
##
##      AIC      BIC    logLik deviance df.resid
##  2128.7   2165.7  -1056.4   2112.7     745
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.91916 -0.90198 -0.01653  0.89486  1.94584
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## conference (Intercept) 0.0000    0.0000
## Residual              0.9684    0.9841
## Number of obs: 753, groups: conference, 4
##
## Fixed effects:
##
##              Estimate Std. Error      df
## (Intercept)   -0.031050   0.045395 753.000000
## format1         0.249944   0.082717 753.000000
## student1       -0.018569   0.078090 753.000000
## gender1         0.126152   0.075351 753.000000
## fleschkincaid_score_scaled  0.047126   0.041553 753.000000
## gender1:fleschkincaid_score_scaled  0.007346   0.082636 753.000000
##
##              t value Pr(>|t|)
## (Intercept)   -0.684   0.4942
## format1         3.022   0.0026 **
## student1       -0.238   0.8121
## gender1         1.674   0.0945 .
## fleschkincaid_score_scaled  1.134   0.2571
## gender1:fleschkincaid_score_scaled  0.089   0.9292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) formt1 stdnt1 gendr1 flsc__
## format1   -0.468
## student1  -0.361  0.090
## gender1    0.274 -0.116  0.024
## flschkncd__ 0.024 -0.136  0.010 -0.022
## gndr1:fls__ -0.082  0.064  0.066 -0.044  0.372
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

Dale-Chall scores:

```

m0 = lmer(Score.mean.norm ~ 1 +
          format + student + gender +
          (1 | conference),
          data = readScores,
          REML = F)

## boundary (singular) fit: see ?isSingular
m1 = update(m0, ~. + dalechall_score_scaled)

## boundary (singular) fit: see ?isSingular
m2 = update(m1, ~. + dalechall_score_scaled:gender)

## boundary (singular) fit: see ?isSingular
anova(m0, m1, m2)

## Data: readScores
## Models:
## m0: Score.mean.norm ~ 1 + format + student + gender + (1 | conference)
## m1: Score.mean.norm ~ format + student + gender + (1 | conference) +
##     dalechall_score_scaled
## m2: Score.mean.norm ~ format + student + gender + (1 | conference) +
##     dalechall_score_scaled + gender:dalechall_score_scaled
##      Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0   6 2126.2 2153.9 -1057.1  2114.2
## m1   7 2128.1 2160.5 -1057.1  2114.1 9e-04      1    0.9763
## m2   8 2130.1 2167.1 -1057.1  2114.1 3e-04      1    0.9869

summary(m2)

## Linear mixed model fit by maximum likelihood . t-tests use
## Satterthwaite's method [lmerModLmerTest]
## Formula: Score.mean.norm ~ format + student + gender + (1 | conference) +
##         dalechall_score_scaled + gender:dalechall_score_scaled
## Data: readScores
##
##      AIC      BIC  logLik deviance df.resid
## 2130.1  2167.1 -1057.1  2114.1      745
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.89133 -0.90206 -0.03695  0.90787  1.89630
##
## Random effects:
##  Groups      Name      Variance Std.Dev.
## conference (Intercept) 0.0000  0.000
## Residual              0.9702  0.985
## Number of obs: 753, groups: conference, 4
##
## Fixed effects:
##
##              Estimate Std. Error      df t value
## (Intercept) -0.033822  0.045339 753.000000  -0.746
## format1      0.265884  0.083397 753.000000   3.188
## student1     -0.017645  0.078048 753.000000  -0.226
## gender1      0.127017  0.075463 753.000000   1.683

```

```

## dalechall_score_scaled      0.001215    0.038208 753.000000    0.032
## gender1:dalechall_score_scaled 0.001233    0.074958 753.000000    0.016
##                               Pr(>|t|)
## (Intercept)                 0.45591
## format1                     0.00149 **
## student1                    0.82120
## gender1                     0.09276 .
## dalechall_score_scaled      0.97464
## gender1:dalechall_score_scaled 0.98689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) formt1 stdnt1 gendr1 dlch__
## format1      -0.466
## student1     -0.359  0.091
## gender1       0.274 -0.121  0.024
## dlchl1_scr_  0.061 -0.202 -0.026  0.017
## gndr1:dlc__ -0.037  0.052  0.026 -0.050  0.147
## convergence code: 0
## boundary (singular) fit: see ?isSingular

```

No interactions.

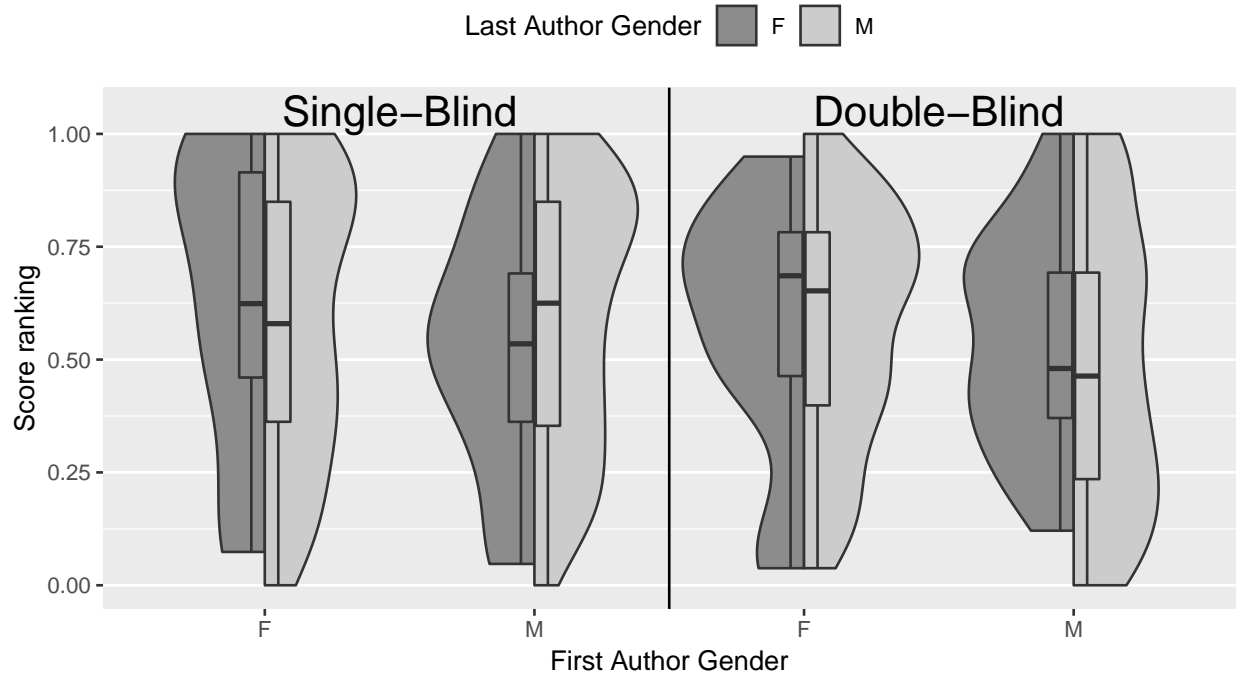


Figure 1: Distributions of review scores by first and last author gender.

Influence of last author

This study considered first authors, but future research could explore the effect of supervising authors and institutions. The data in this study is not ideal for exploring this, since the number of papers with multiple authors varies between conferences and there are many non-independencies. The raw data is not made available here because the combination of factors make cases identifiable.

We investigated whether the review scores were biased by combinations of first author gender, last author gender and review type (mixed effects model with a random intercept for each conference). We note that the biggest change is for male-male authors from E10 (single-blind) to E11 (double-blind), which would be consistent with a gender bias being neutralised by double-blind review. However, statistically, there was only a significant main effect of format.

Here are the distributions of review scores by first and last author gender:

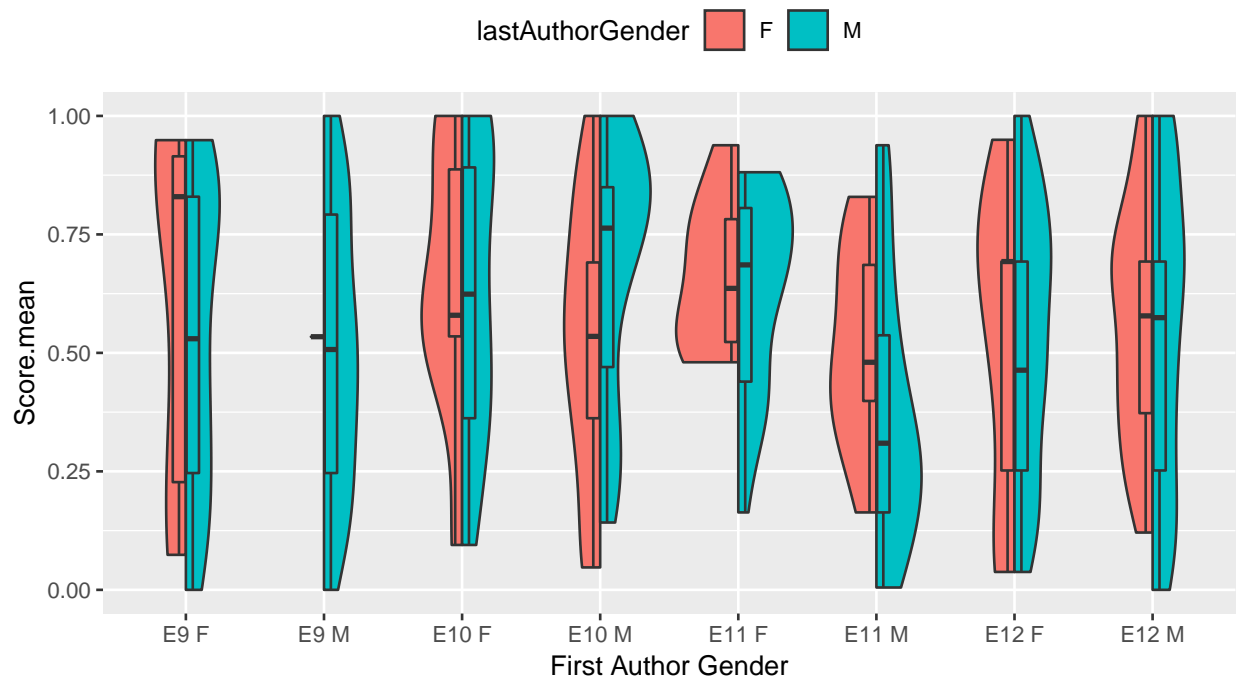


Figure 2: Distributions of review scores by first and last author gender.

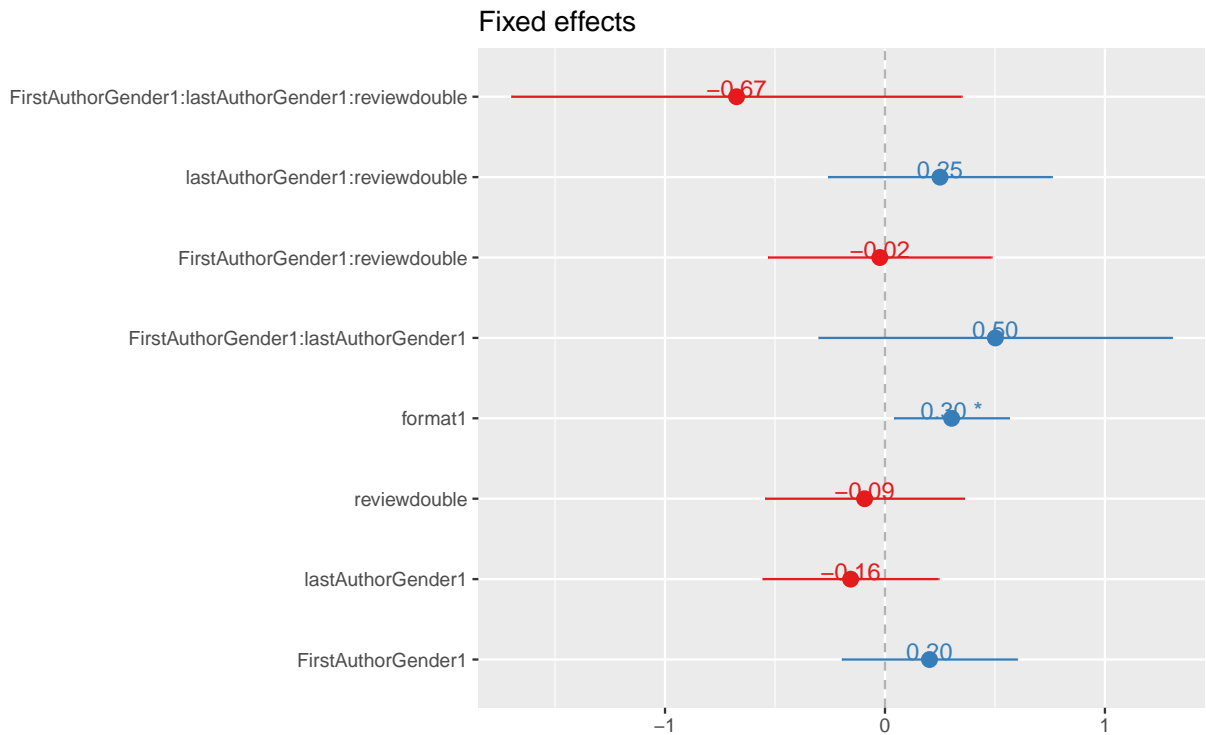


Figure 3: Coefficients and confidence intervals for effects predicting review ranks.