The impact of double blind reviewing at EvoLang 12: statistics

Contents

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)
```

```
# 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
```

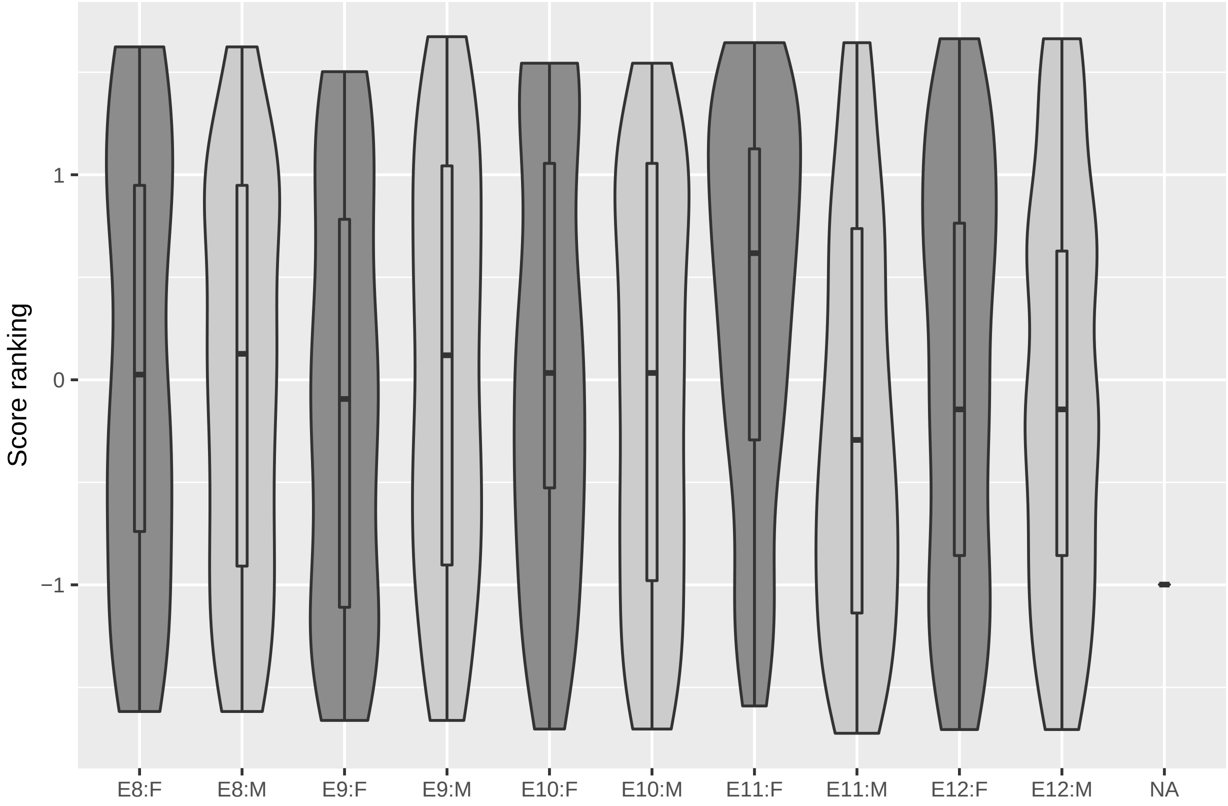
```
## , ,  = Non-Student
##
##
##      E8 E9 E10 E11 E12
##    F  0 34  55  41  54
##    M  0 85  94  77  93
##
## , ,  = Student
##
##
```

```
##      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.

```r
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
```

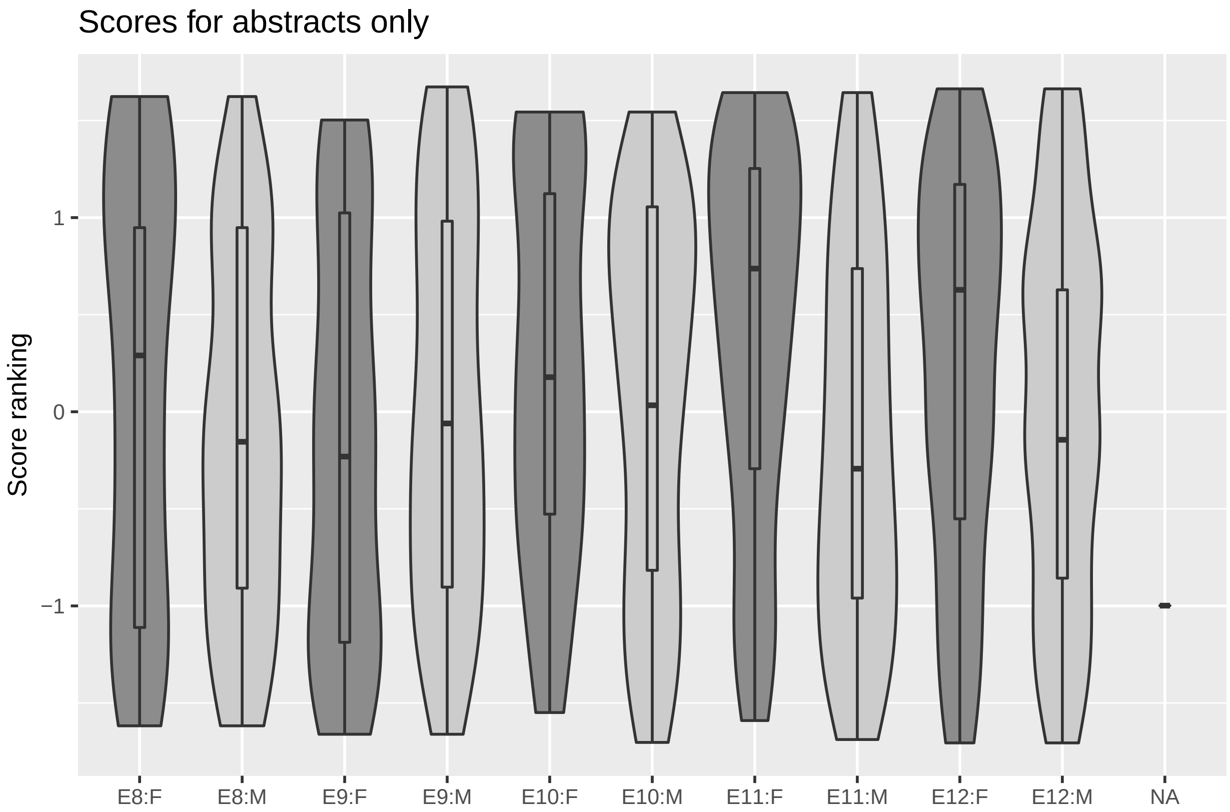


```r
pdf("../results/Results_Gender_3conf.pdf", width = 12, height= 6)
p2
dev.off()
```

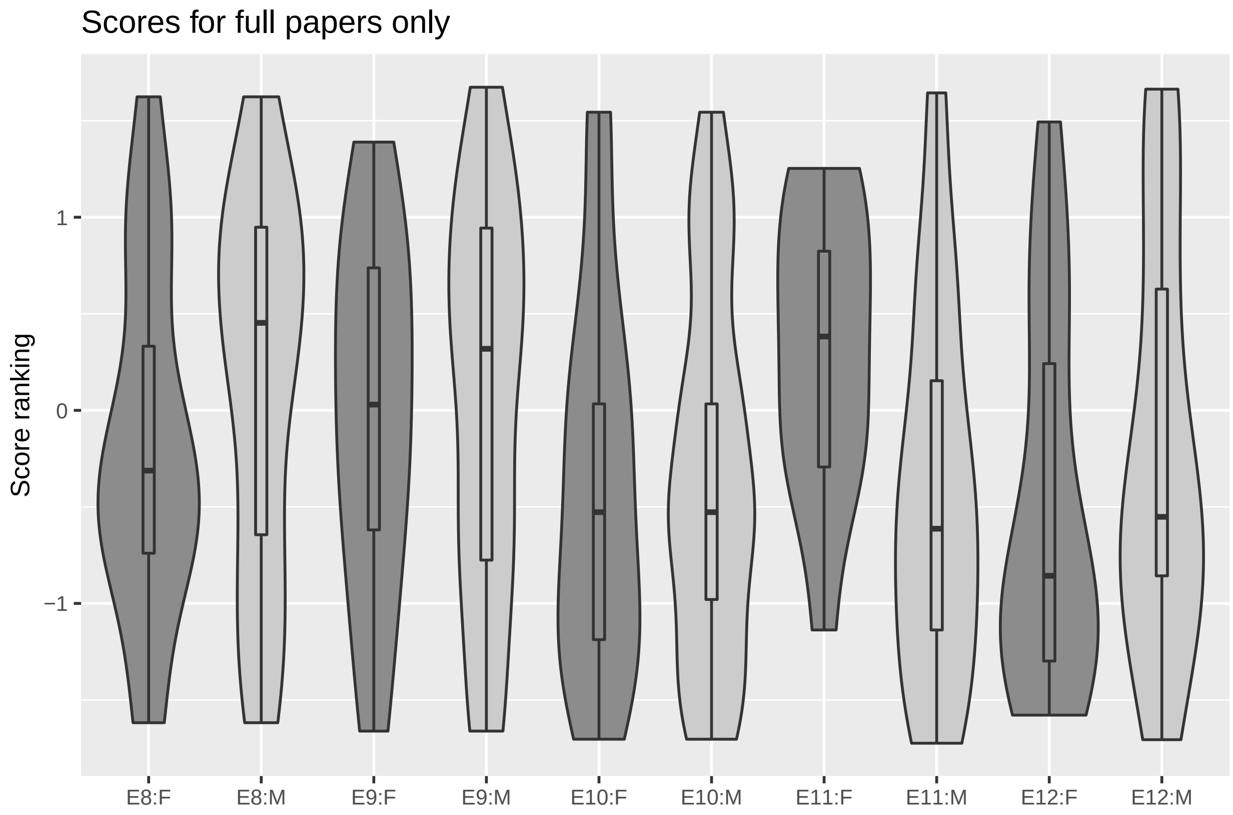
```
## pdf
##   2
```

```r
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
```

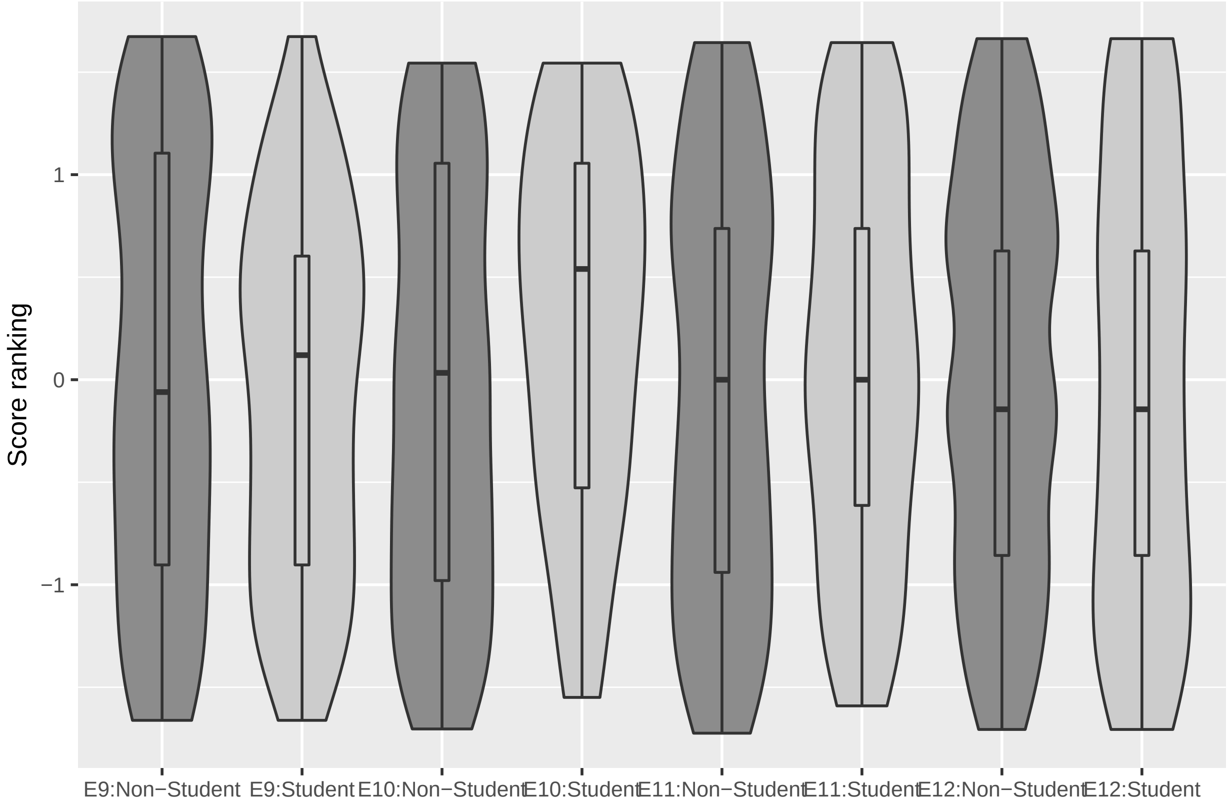


```
p2Paper <- ggplot(allData[allData$format=="Paper",],
             aes((conference):(FirstAuthorGender), Score.mean,
                 fill=FirstAuthorGender))

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
```

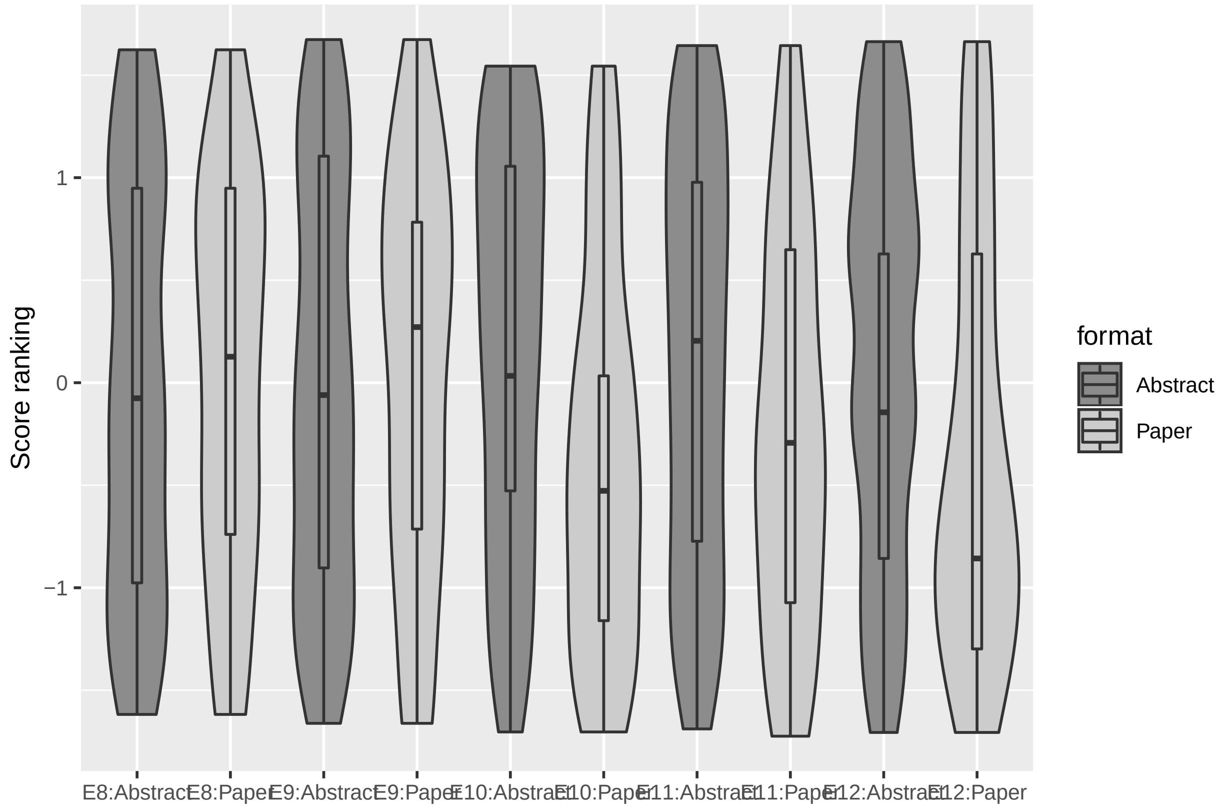


```r
pdf("../results/Results_Student_3conf.pdf", width = 12, height= 6)
p
dev.off()
```

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

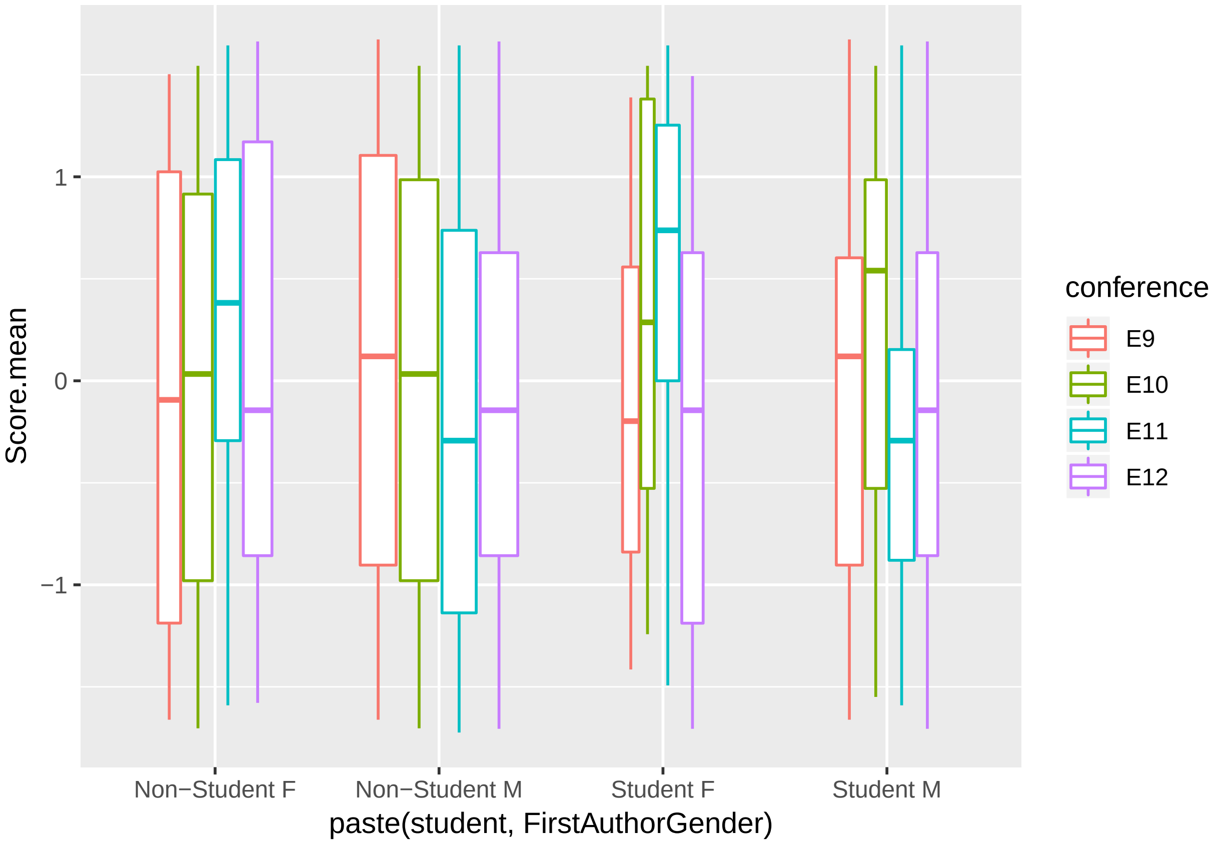
Format:

```r
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
```

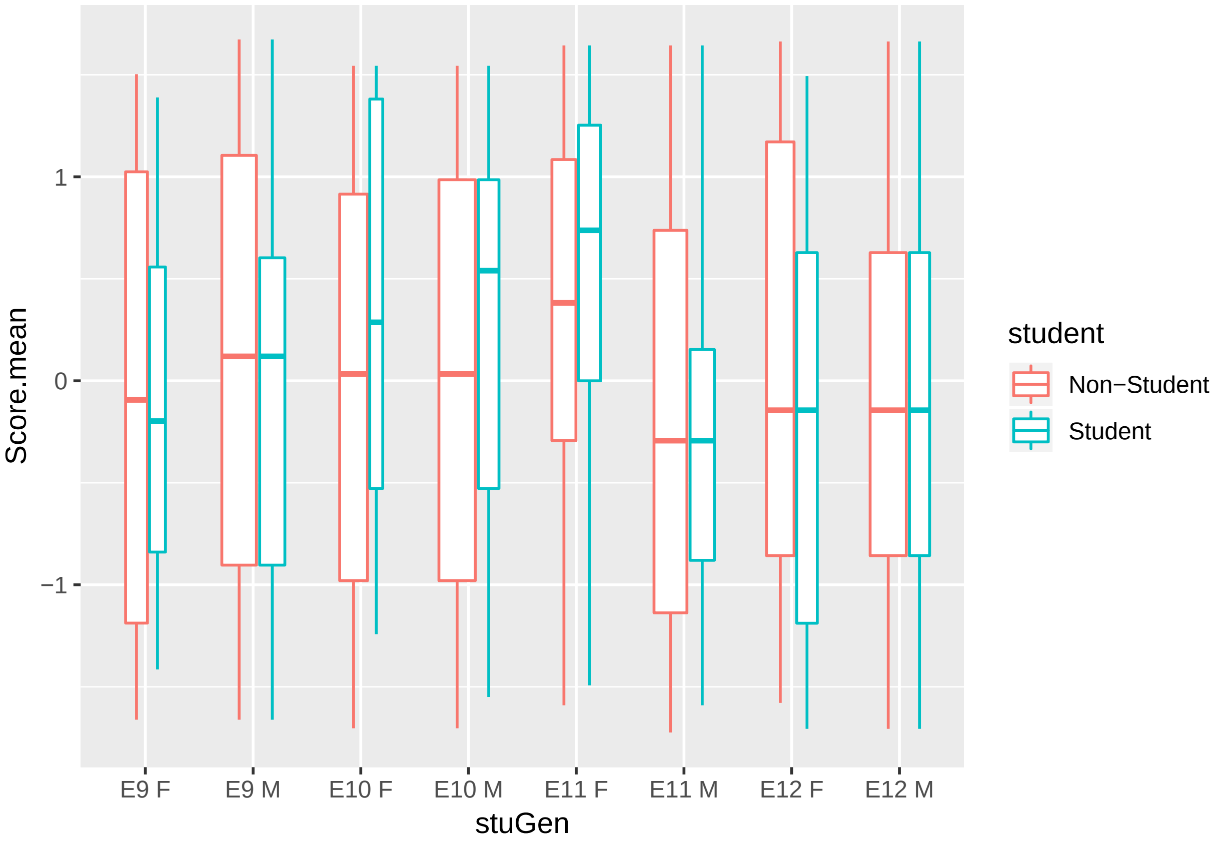


Combined student and gender:

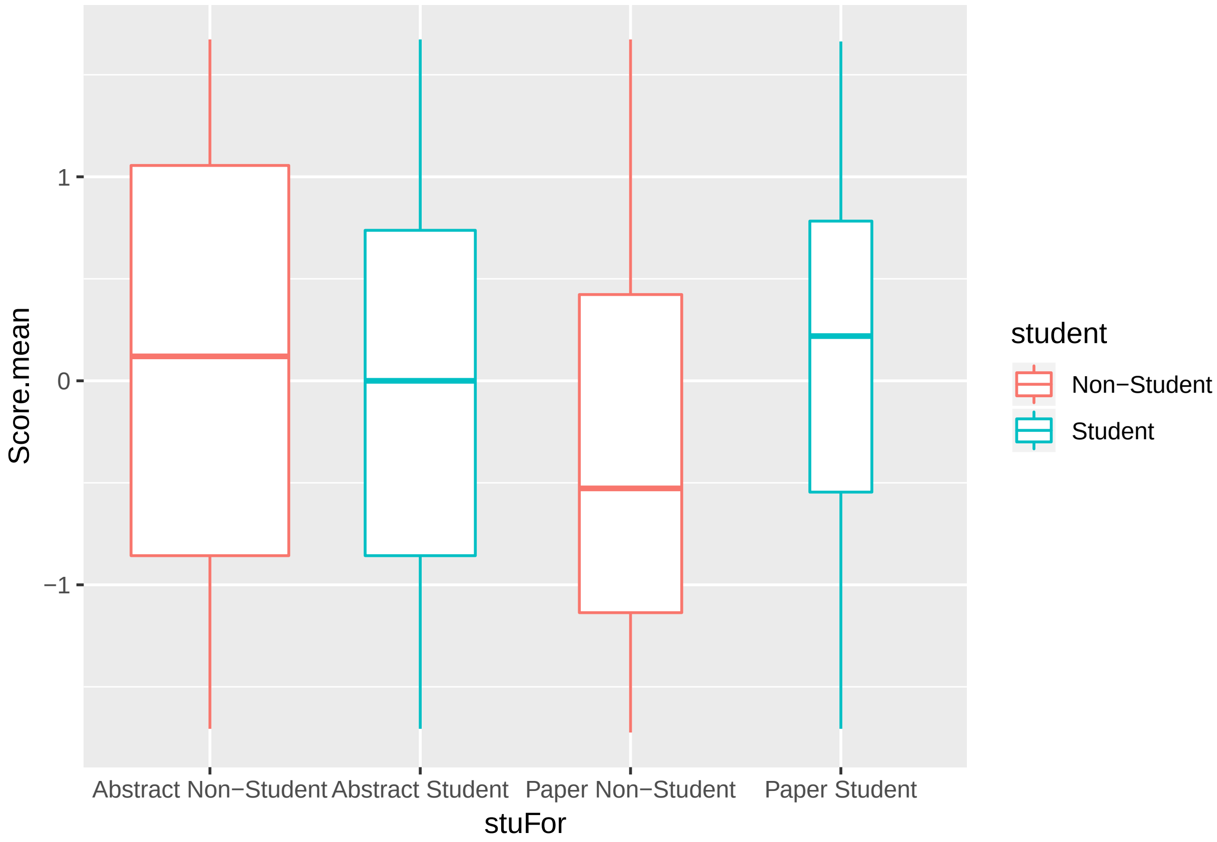
```
ggplot(allData[allData$conference!="E8",],
       aes(y=Score.mean,x=paste(student,FirstAuthorGender),colour=conference))+ geom_boxplot(varwidth =
```



```
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

```r
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
```

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 permissable, 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=10000000)),
  REML = T
)

summary(m0)
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## 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                           0.146719   0.166438
## 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)                                   2.900000  -0.087   0.9367
## FirstAuthorGender1                            2.600000   0.882   0.4519
## review1                                       2.900000  -0.738   0.5163
## student1                                      2.900000  -1.428   0.2528
## format1                                       3.400000   1.269   0.2845
## FirstAuthorGender1:review1                    2.600000   0.771   0.5046
## FirstAuthorGender1:student1                 674.800000  -1.101   0.2714
## review1:student1                              2.900000   0.762   0.5040
## FirstAuthorGender1:format1                  749.300000   0.467   0.6405
## review1:format1                               3.400000   1.178   0.3148
## student1:format1                            615.300000   3.278   0.0011 **
## FirstAuthorGender1:review1:student1         674.800000   0.186   0.8526
## FirstAuthorGender1:review1:format1          749.300000   0.474   0.6357
## FirstAuthorGender1:student1:format1         719.900000   0.663   0.5077
## review1:student1:format1                    615.300000  -1.434   0.1521
## FirstAuthorGender1:review1:student1:format1 719.900000   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
```

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=10000000)),
  REML = T
)

summary(m0)
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## 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   6.10000  -1.069
## FirstAuthorGender1                   0.11629    0.11771   3.90000   0.988
## review1                             -0.02834    0.09293   6.10000  -0.305
## format1                              0.26421    0.12964   3.40000   2.038
## FirstAuthorGender1:review1           0.29146    0.23542   3.90000   1.238
## FirstAuthorGender1:format1           0.21076    0.15756 904.20000   1.338
## review1:format1                      0.17404    0.25928   3.40000   0.671
## FirstAuthorGender1:review1:format1  -0.05484    0.31512 904.20000  -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
```

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

Permutation test

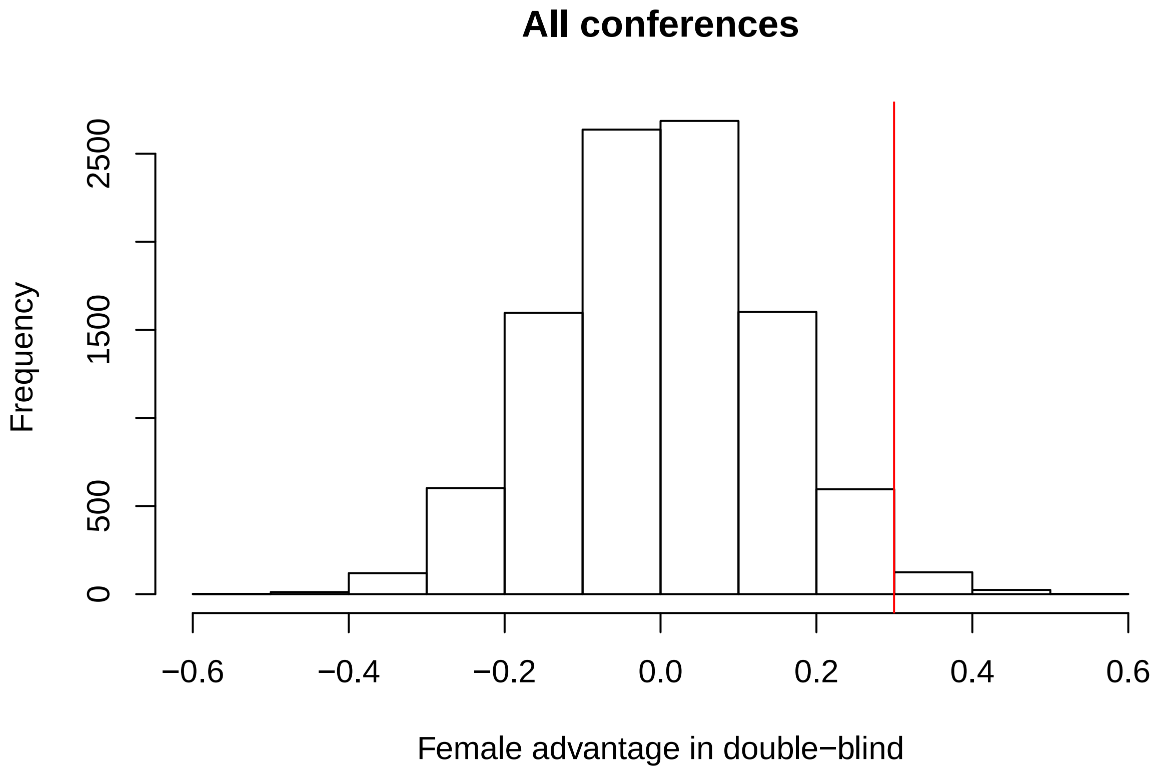
The distributions of score means are not very normal within conferences. We run a permutation test to address this. We calculate the average difference between single blind and double blind scores for males (dM) and for females (dF). Then we calculate dF - dM. A value > 0 means females scores increase 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")
```

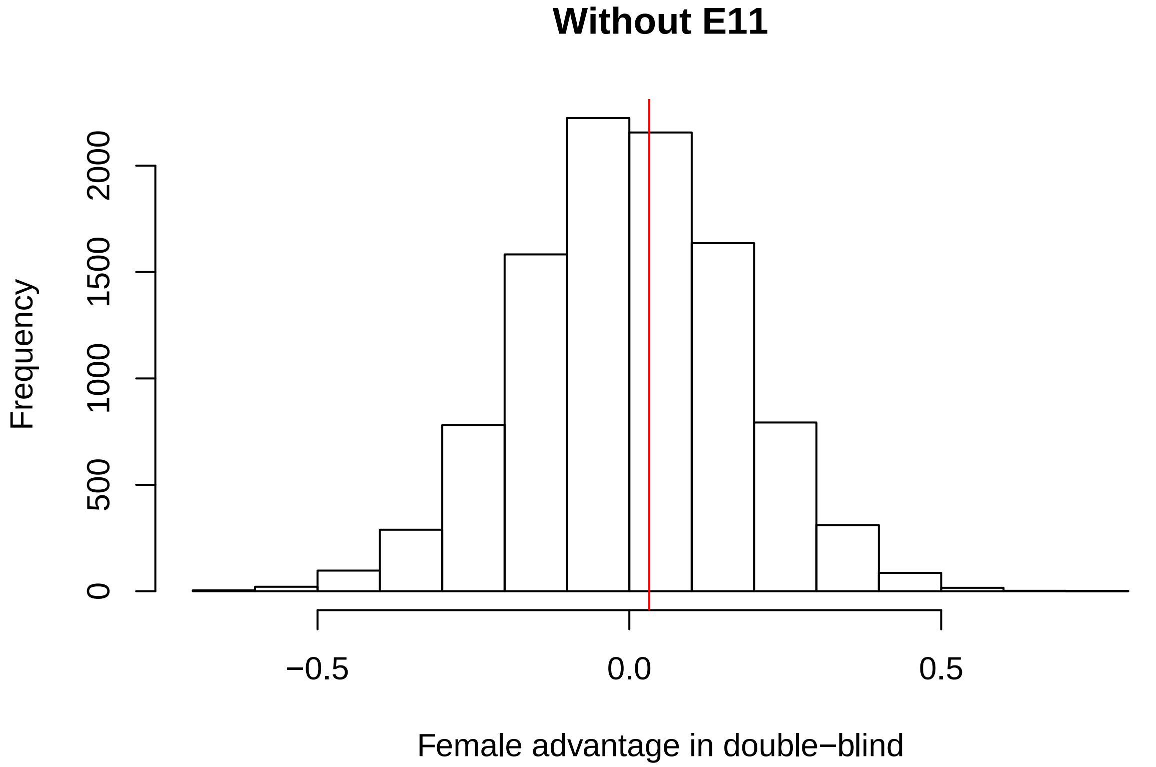
```
## [1] "p= 0.015 , z= 2.17305837569821"
```



Permutation test without E11 data:

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

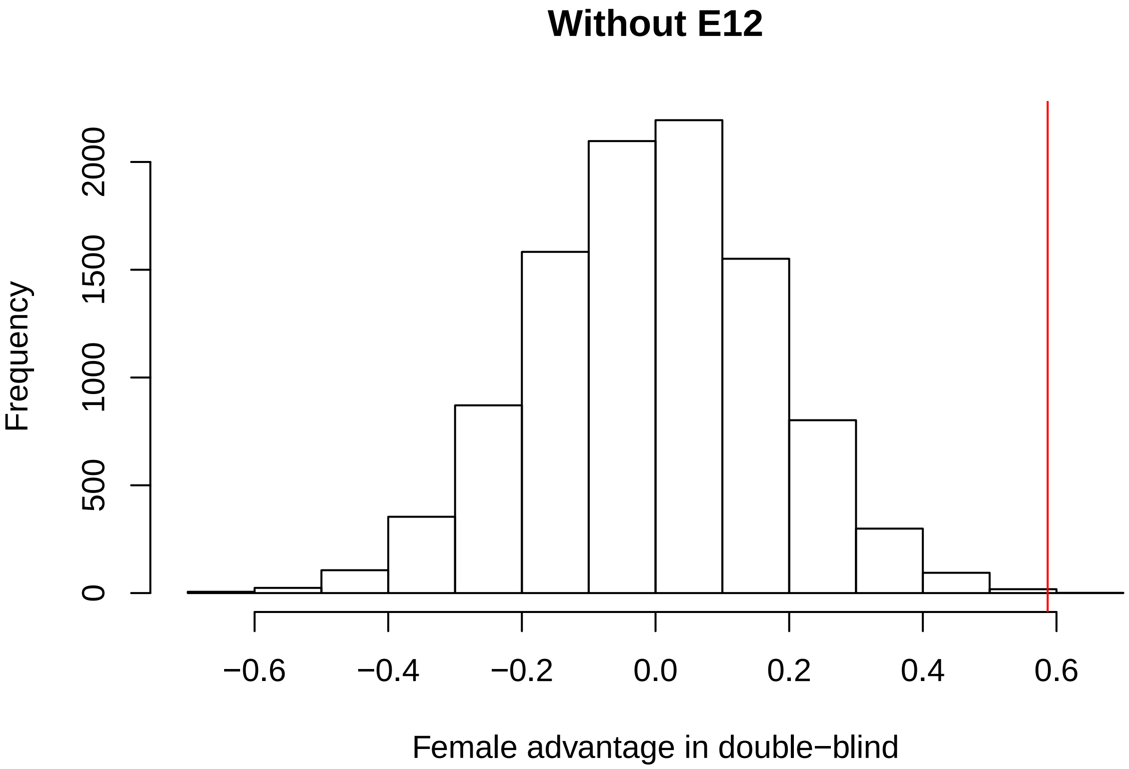
```
## [1] "p= 0.4278 , z= 0.181357792321726"
```



Permutation test without E12 data:

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

```
## [1] "p= 1e-04 , z= 3.33434405436834"
```

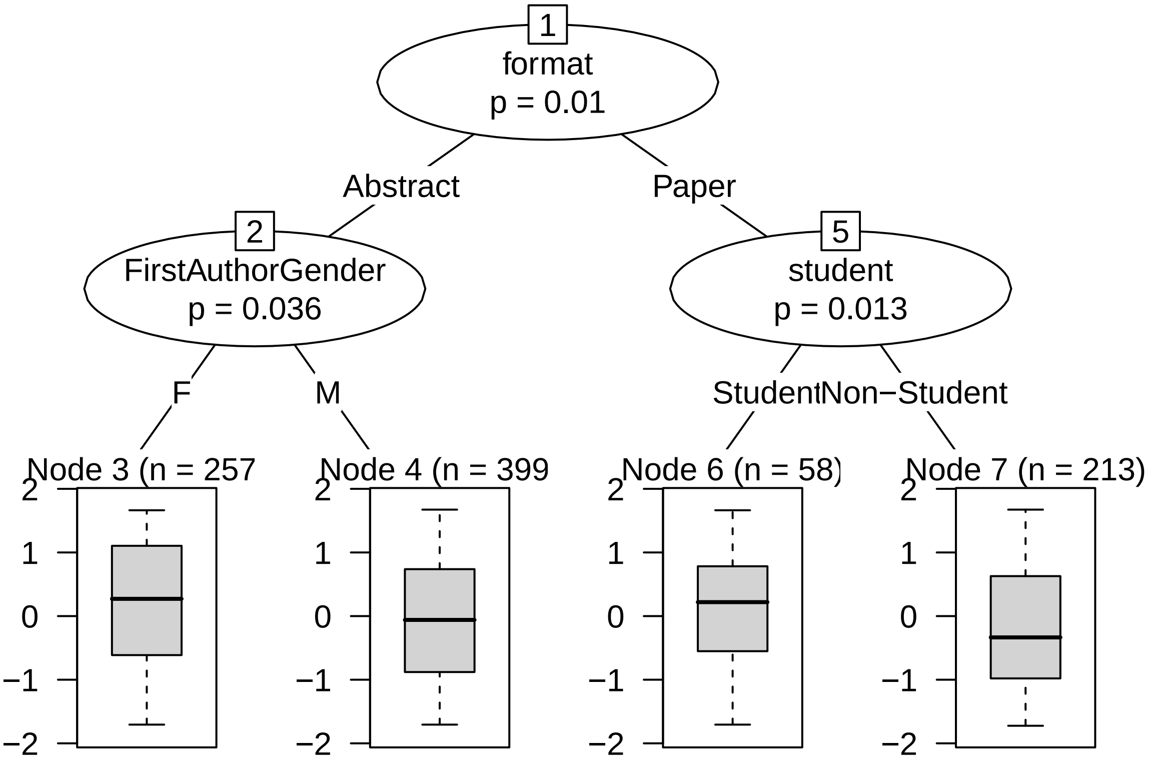


The results are in line with the test above. Across the whole data, females are given higher scores in double-blind, but this is driven by E11 alone.

Decision tree exploration

Construct a decision tree, attempting to predict review socres 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
```

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

```
## 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 centrered.
- `*_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-Kinkaid 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            -0.72                0.61             0.61
##                     smog_score dalechall_score
## flesch_score             -0.93           -0.72
## fleschkincaid_score       0.96            0.61
## gunningfog_score          0.99            0.61
## smog_score                1.00            0.64
## dalechall_score           0.64            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] 13.11313
```

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

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  readScores$fleschkincaid_score and readScores$dalechall_score
## t = 23.138, df = 900, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5680890 0.6500779
## sample estimates:
##       cor
## 0.6107177
```

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

```
## [1] -0.4780219
```

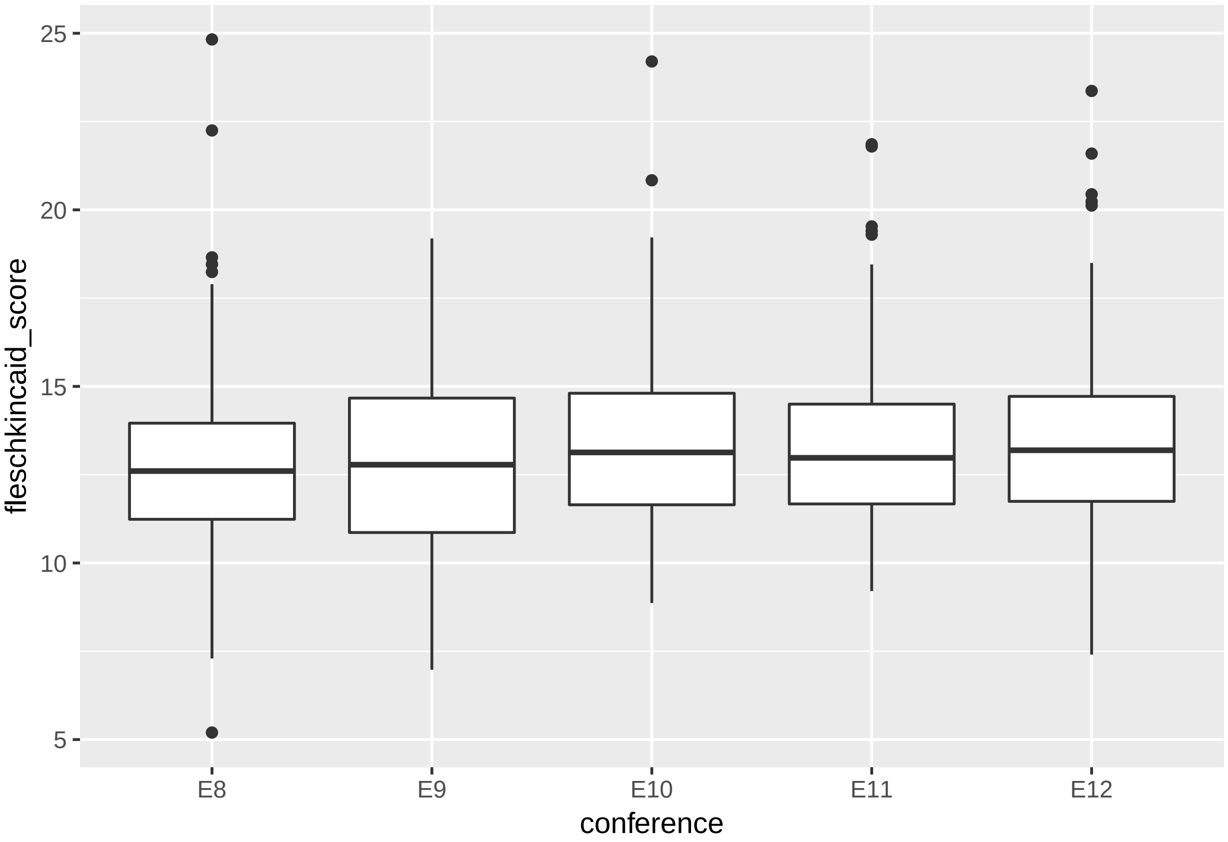
```
sel = readScores$conference=="E12"
mean(readScores[sel & readScores$gender=="M",]$fleschkincaid_score) -
  mean(readScores[sel & readScores$gender=="F",]$fleschkincaid_score)
```

```
## [1] 0.1664231
```

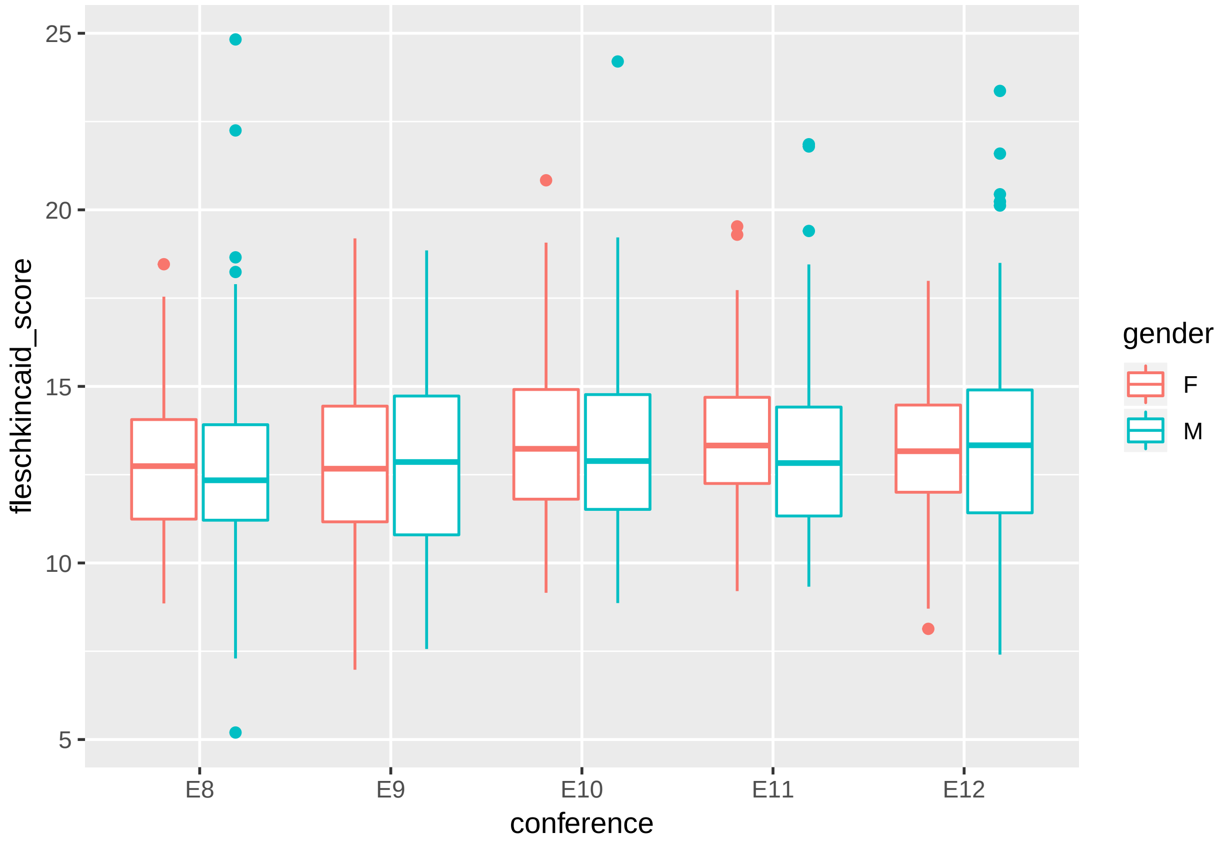
```
meanFK =
  rbind(tapply(readScores$fleschkincaid_score[readScores$gender=="F"],
             readScores$conference[readScores$gender=="F"],mean),
tapply(readScores$fleschkincaid_score[readScores$gender=="M"],
       readScores$conference[readScores$gender=="M"],mean))
sdFK =
  rbind(tapply(readScores$fleschkincaid_score[readScores$gender=="F"],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

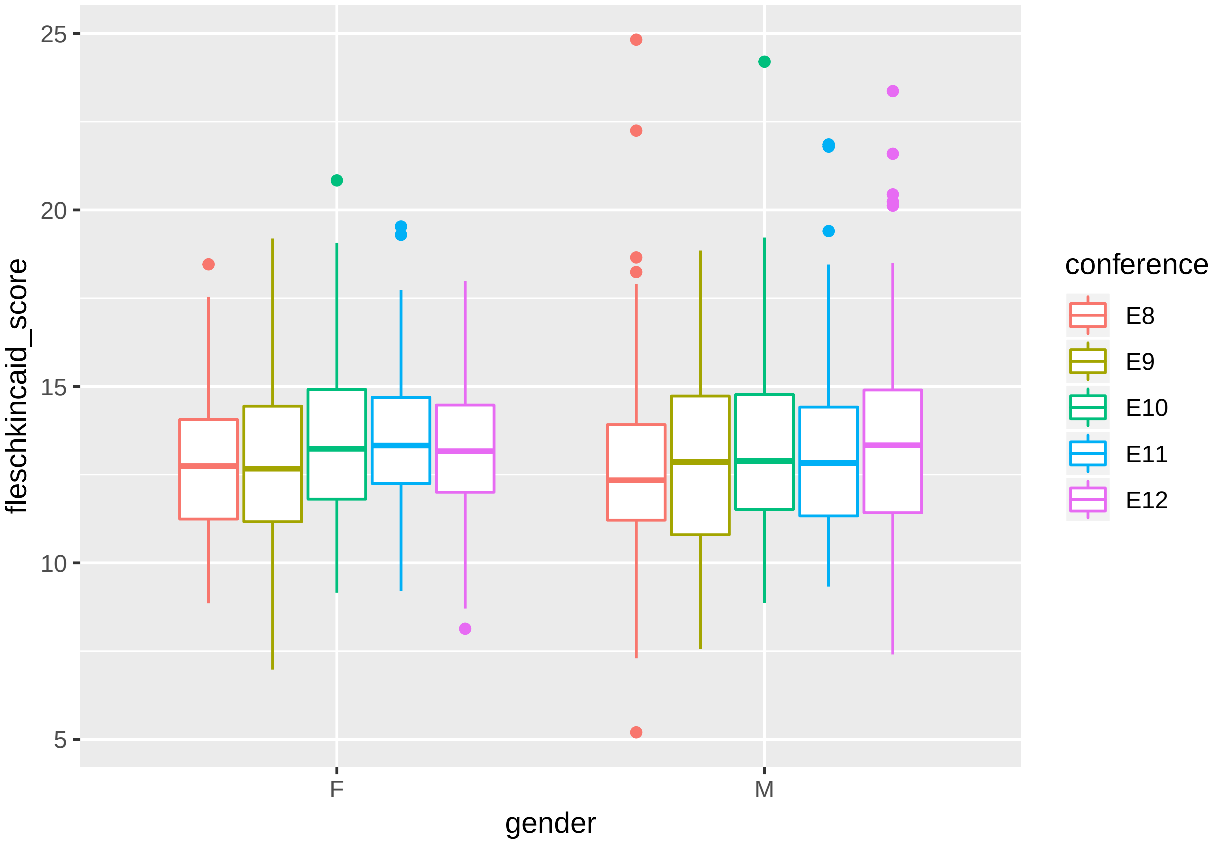
```
readScores$gender2 = "Female"
readScores$gender2[readScores$gender=="M"] = "Male"

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



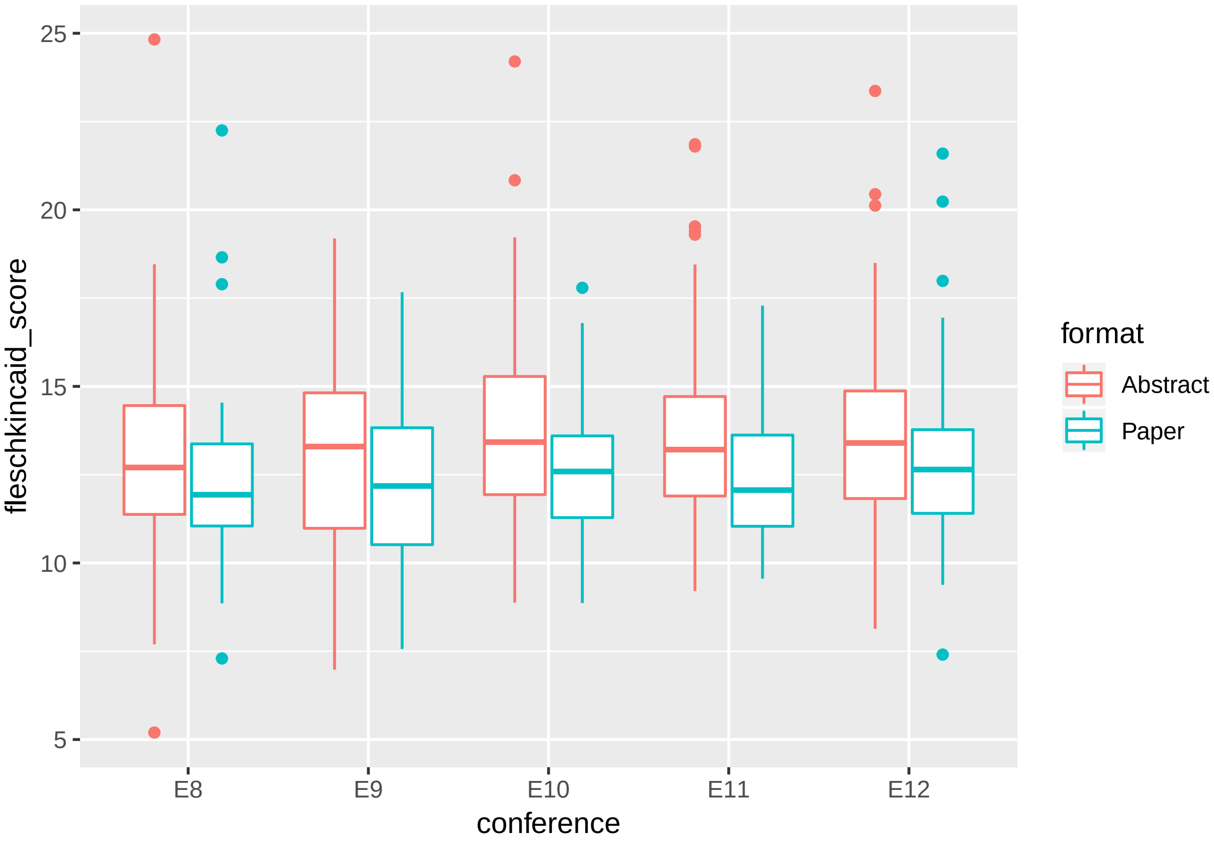
```
ggplot(readScores, aes(y=fleschkincaid_score,x=conference,colour=gender)) + geom_boxplot()
```



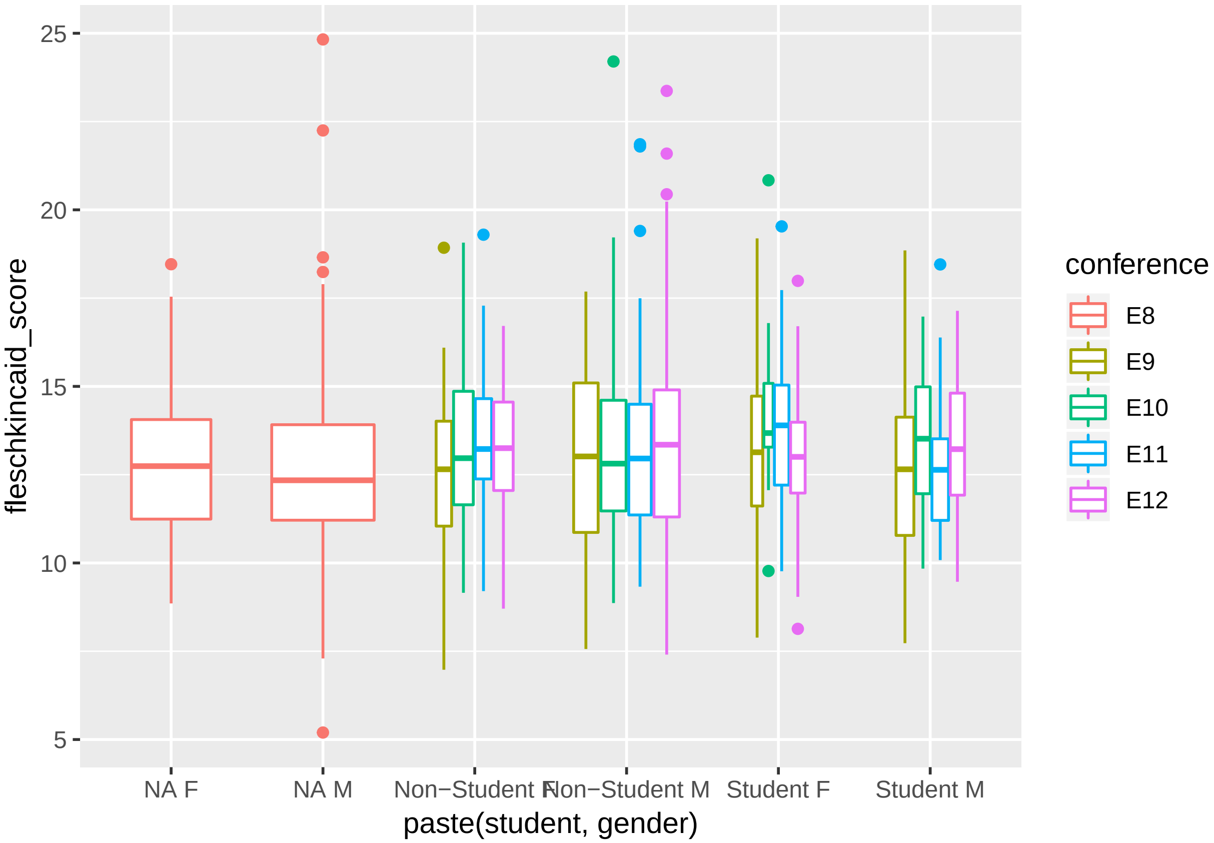
```
ggplot(readScores, aes(y=fleschkincaid_score,x=gender,colour=conference)) + geom_boxplot()
```



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



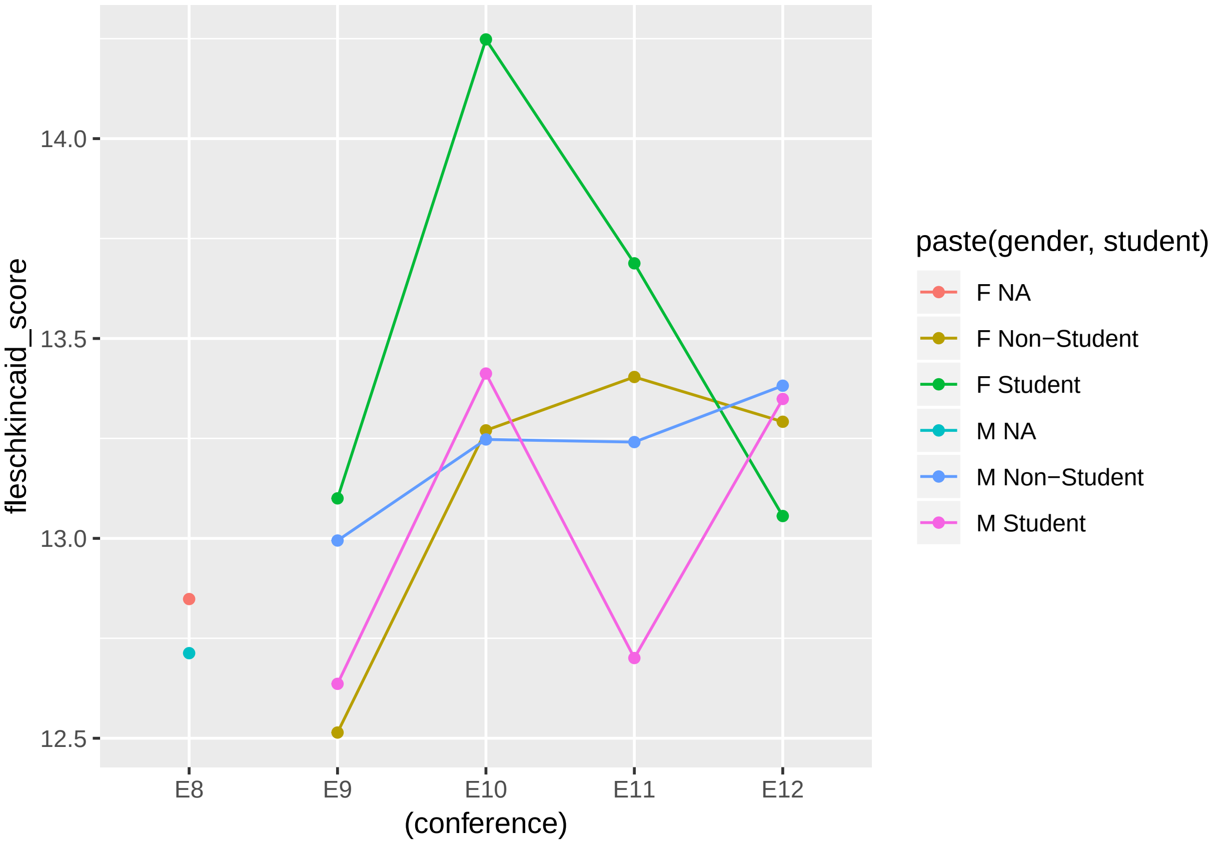
```
ggplot(readScores, aes(y=fleschkincaid_score,x=paste(student,gender),colour=conference))+ geom_boxplot(
```



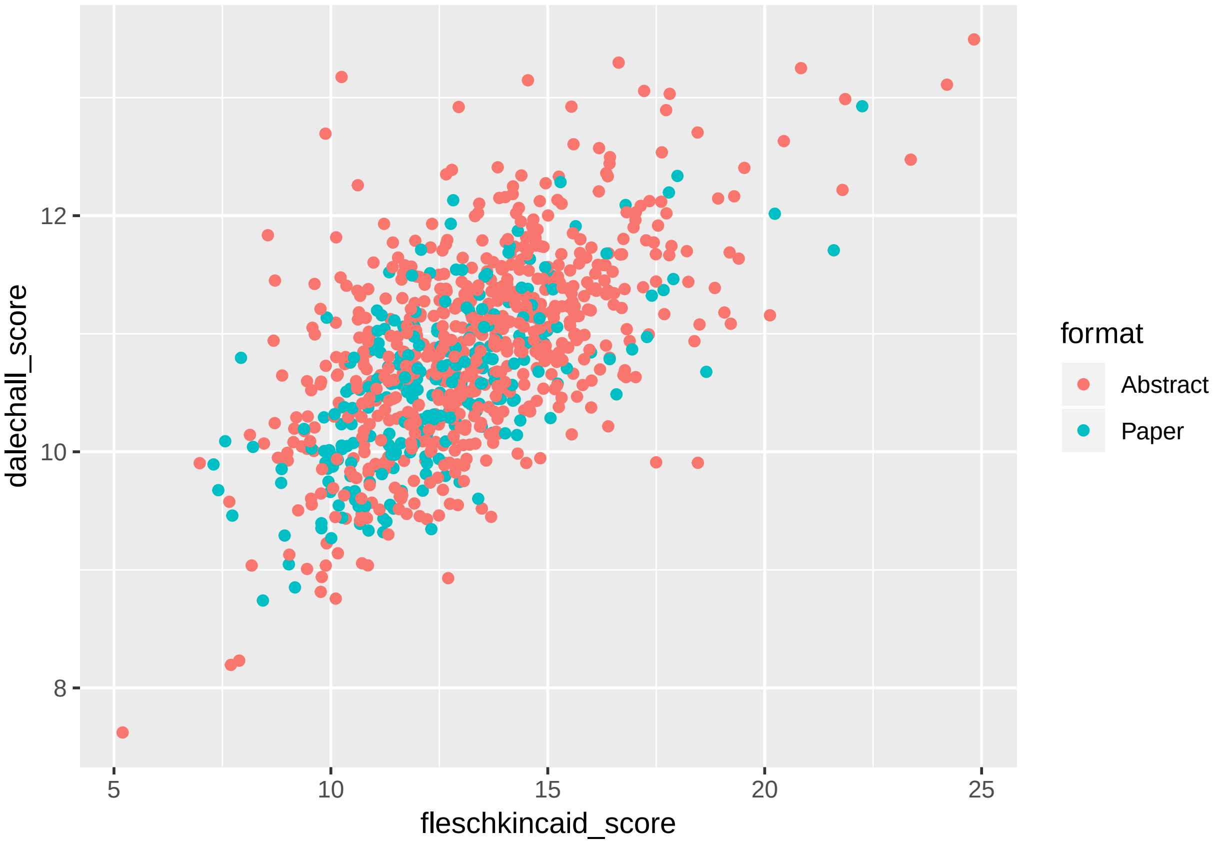
```
fkrs = ggplot(readScores, aes(y=fleschkincaid_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Flesch-Kincaid reading score", x="Gender")

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))
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()
```

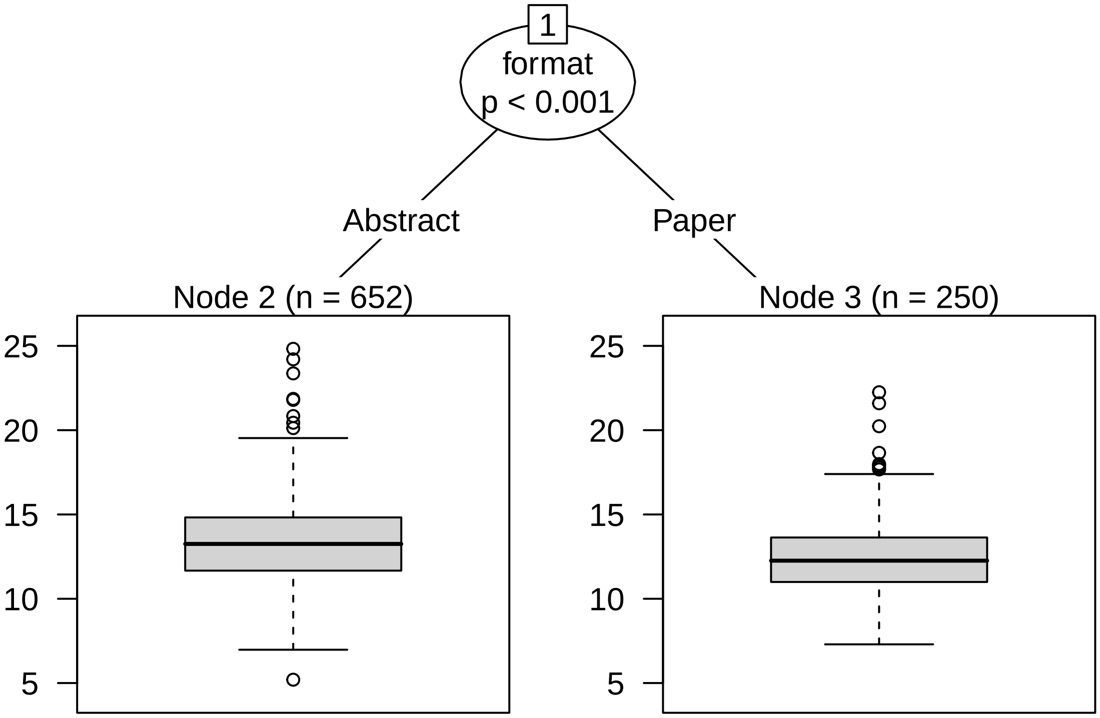


```
fkrs2= ggplot(readScores, aes(y=fleschkincaid_score,
                              x=conference,colour=gender)) +
  geom_boxplot() +
  labs(y="Flesch-Kincaid reading score", x="Conference")

pdf("../results/FleschKincaidReadingScores2.pdf",
    width=6,height=4)
fkrs2
dev.off()
```

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

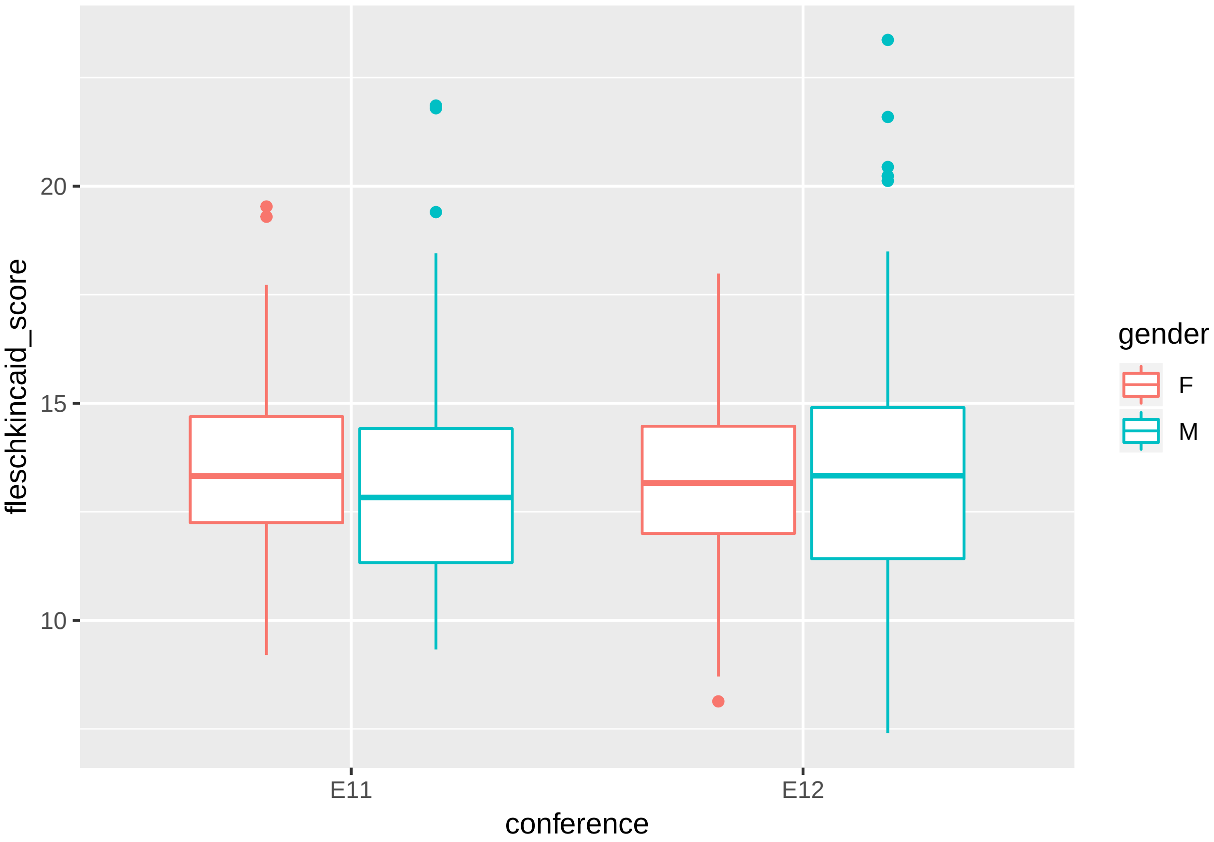
Decision tree

```
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  0.811  0.8108   9.562 0.00214 **
## conference                         1  0.000  0.0001   0.001 0.97103
## student                            1  0.070  0.0701   0.827 0.36372
## gender                             1  0.018  0.0177   0.208 0.64844
## format:conference                  1  0.111  0.1112   1.311 0.25294
## format:student                     1  0.062  0.0624   0.736 0.39151
## conference:student                 1  0.015  0.0148   0.175 0.67637
## format:gender                      1  0.113  0.1130   1.333 0.24905
## conference:gender                  1  0.057  0.0567   0.669 0.41396
## student:gender                     1  0.036  0.0355   0.419 0.51784
## format:conference:student          1  0.112  0.1124   1.325 0.25039
## format:conference:gender           1  0.002  0.0017   0.020 0.88807
## format:student:gender              1  0.004  0.0038   0.045 0.83230
## conference:student:gender          1  0.099  0.0986   1.162 0.28168
## format:conference:student:gender   1  0.219  0.2190   2.583 0.10889
## Residuals                        368 31.205  0.0848
## ---
## 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",])
summary(m0)
```

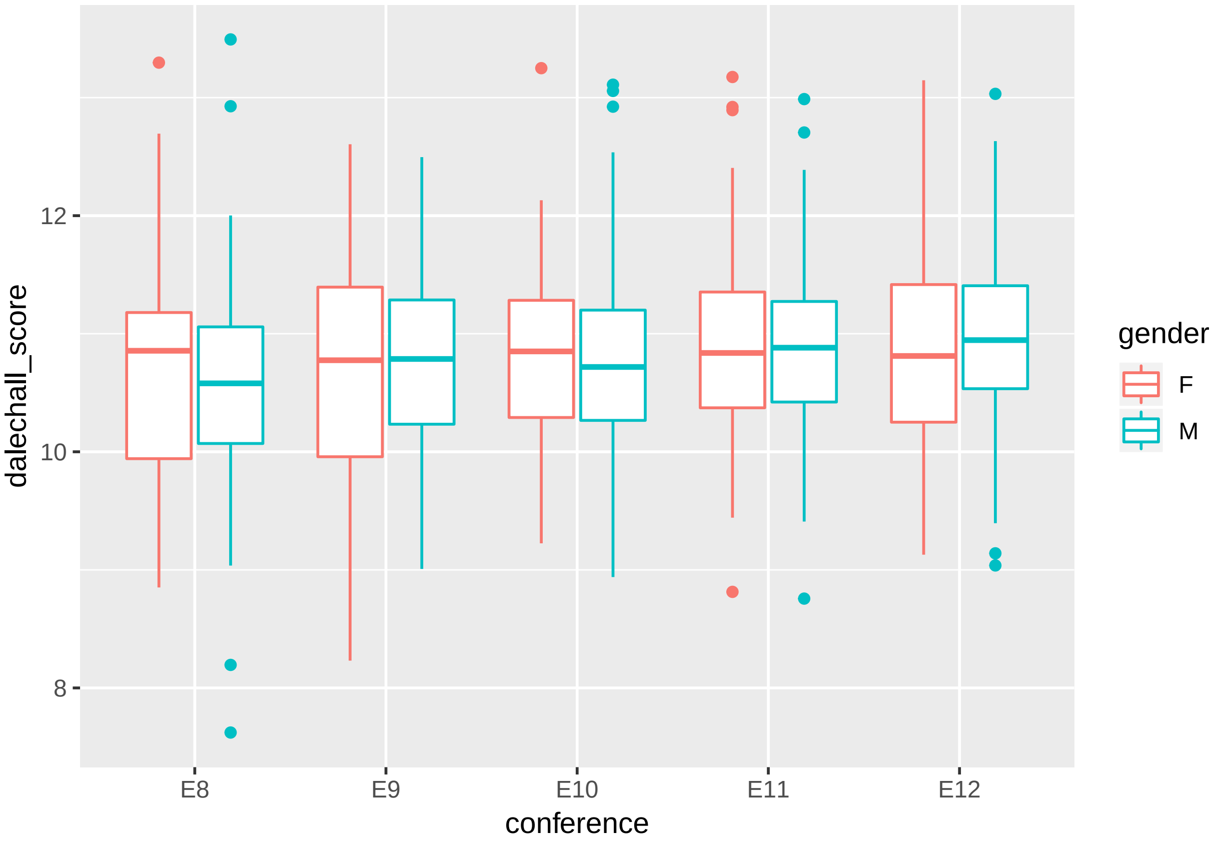
```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula: fleschkincaid_score_scaled ~ 1 + (format + student + gender +
##     review)^2 + time + (1 + format + student + gender | conference)
##    Data: readScores[readScores$conference != "E8", ]
##
## REML criterion at convergence: 2053.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4417 -0.7021 -0.0512  0.5907  4.3979
##
## Random effects:
##  Groups     Name        Variance  Std.Dev. Corr
##  conference (Intercept) 0.0055514 0.07451
##             format1     0.0054120 0.07357   1.00
##             student1    0.0003985 0.01996  -1.00 -1.00
##             gender1     0.0071636 0.08464   1.00  1.00 -1.00
##  Residual               0.8678717 0.93160
## Number of obs: 753, groups:  conference, 4
##
## Fixed effects:
##                        Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)            -0.20192    0.15846   3.10000  -1.274   0.2900
## format1                 0.28655    0.13768   8.30000   2.081   0.0696 .
## student1                0.07745    0.11835  65.40000   0.654   0.5151
## gender1                 0.13950    0.12964   5.90000   1.076   0.3242
## reviewSingle            0.20255    0.21439   3.00000   0.945   0.4138
## time                    0.11253    0.09008   3.00000   1.249   0.3008
## format1:student1        0.02661    0.17604 738.40000   0.151   0.8799
## format1:gender1        -0.26543    0.16999 736.80000  -1.561   0.1188
## format1:reviewSingle    0.02550    0.17199   5.10000   0.148   0.8878
## student1:gender1       -0.25645    0.15545 726.70000  -1.650   0.0994 .
## student1:reviewSingle  -0.15616    0.15276  43.90000  -1.022   0.3123
## gender1:reviewSingle   -0.04597    0.16817   4.10000  -0.273   0.7977
## ---
## 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      -0.084
## student1     -0.138  0.179
## gender1      0.281 -0.013 -0.007
## reviewSingl -0.915  0.030  0.063 -0.189
## time        -0.847 -0.034 -0.065 -0.034  0.843
## frmt1:stdn1  0.073 -0.307 -0.477 -0.003  0.000  0.037
## frmt1:gndr1 -0.144  0.240 -0.018 -0.419  0.075  0.061  0.030
## frmt1:rvwSn  0.032 -0.679 -0.035 -0.072 -0.006  0.017  0.015    0.003
## stdnt1:gnd1 -0.015  0.015  0.125 -0.199 -0.029  0.025 -0.079    0.059
## stdnt1:rvwS  0.133 -0.067 -0.635 -0.027 -0.205 -0.006  0.119    0.001
## gndr1:rvwSn -0.169 -0.039 -0.009 -0.640  0.270 -0.011 -0.003    0.099
##             frm1:S std1:1 std1:S
## format1
## student1
## gender1
## reviewSingl
## time
## frmt1:stdn1
## frmt1:gndr1
## frmt1:rvwSn
## stdnt1:gnd1  0.007
## stdnt1:rvwS  0.018  0.125
## gndr1:rvwSn  0.132 -0.072 -0.062
```

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

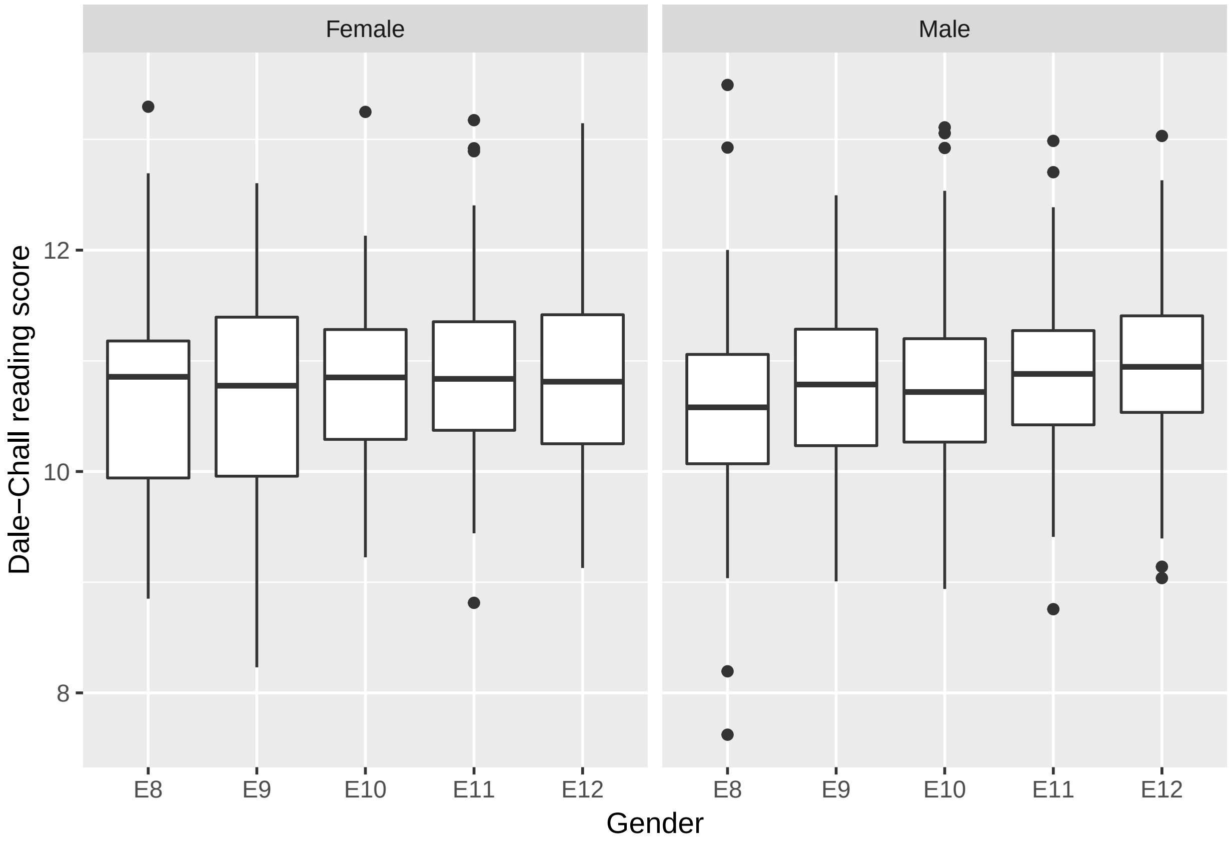
Dale-Chall scale

Plots

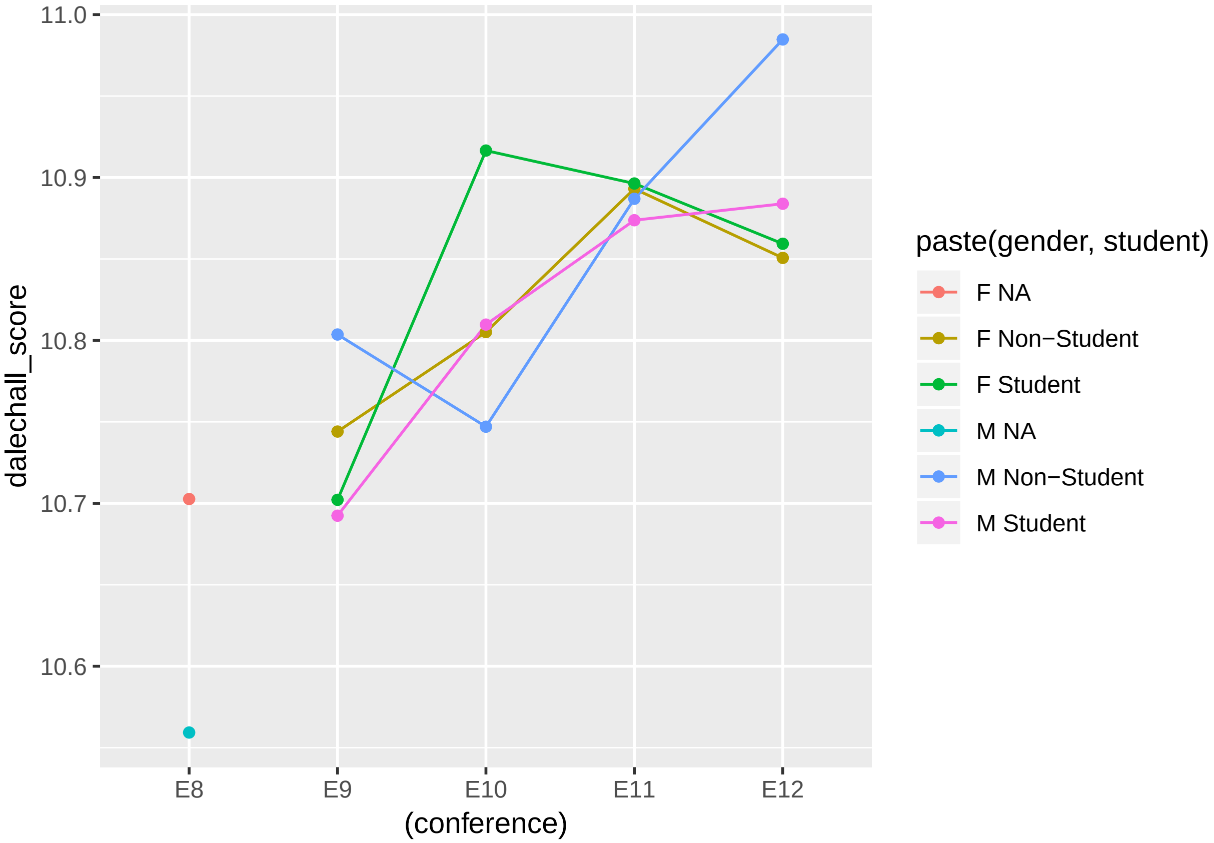
```
ggplot(readScores, aes(y=dalechall_score,x=conference,colour=gender)) + geom_boxplot()
```



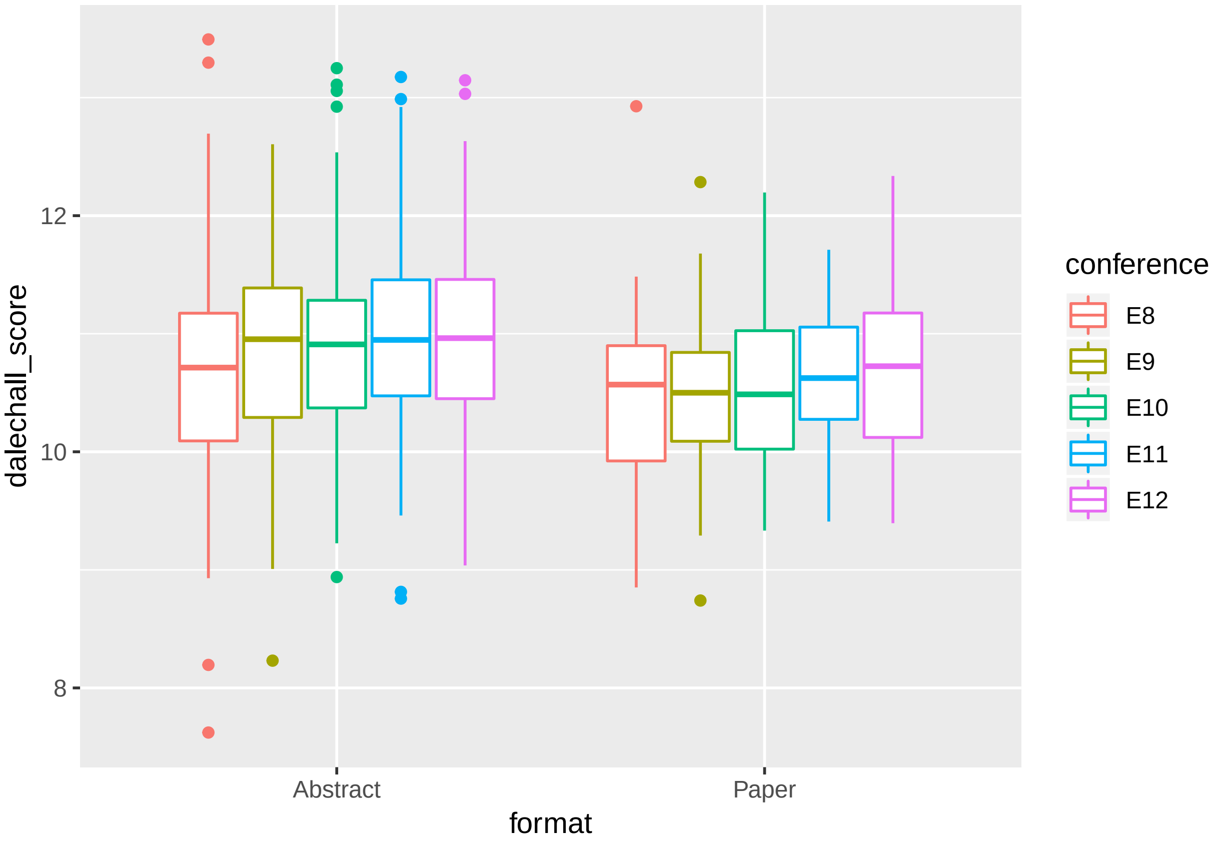
```
ggplot(readScores, aes(y=dalechall_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Dale-Chall reading score", x="Gender")
```



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

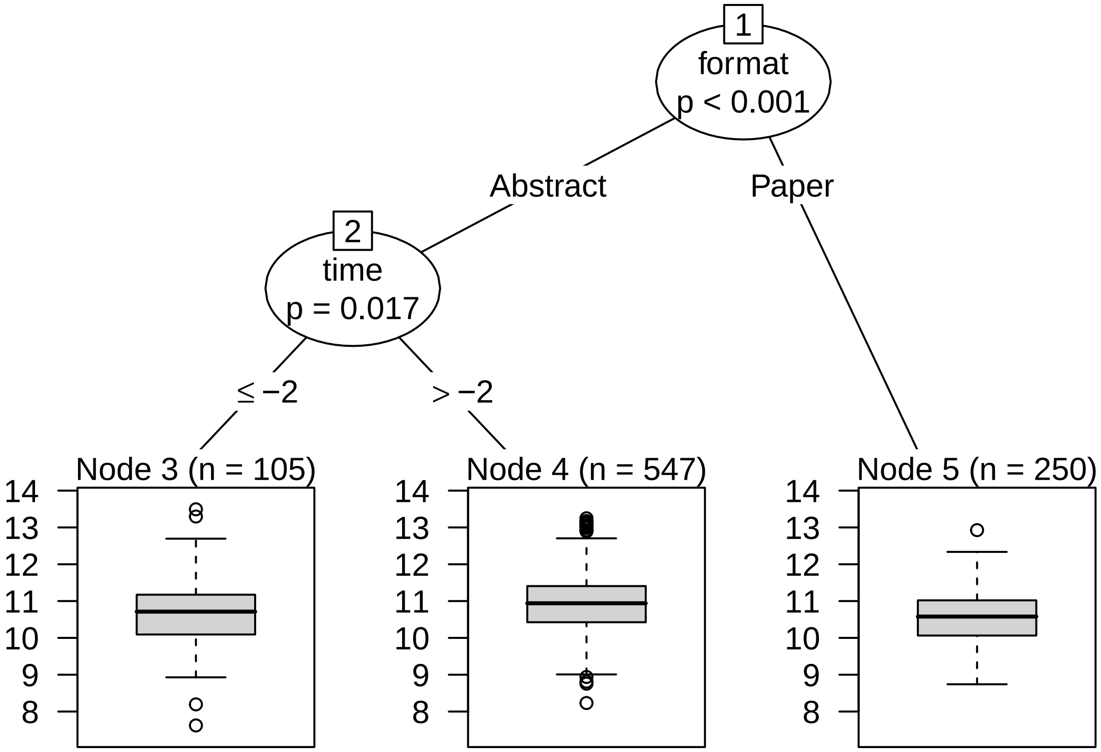


```
ggplot(readScores, aes(y=dalechall_score,x=format,colour=conference)) + geom_boxplot()
```


dalechall_score
12
10
8
Abstract
Paper
format
conference
E8
E9
E10
E11
E12

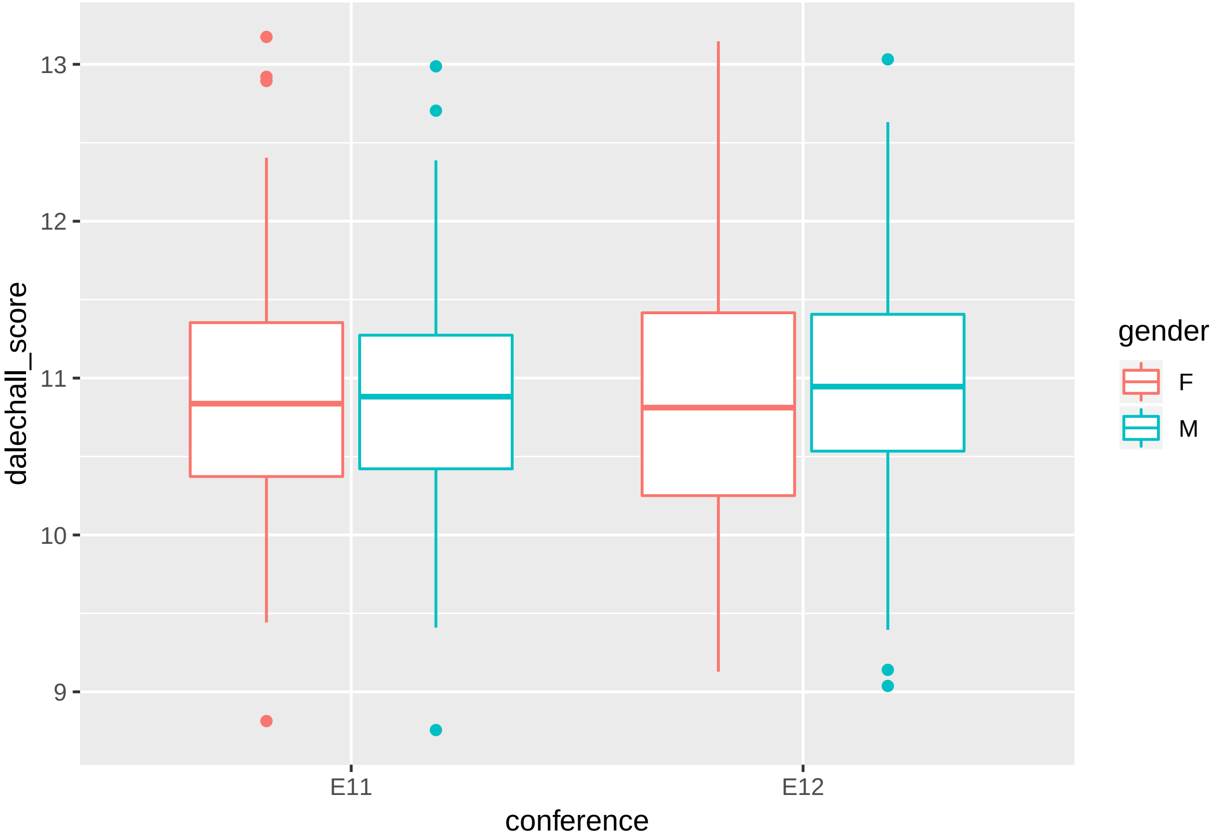
Decision tree:

```
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   1.72  1.7179  12.240 0.000525 ***
## conference                        1   0.01  0.0060   0.043 0.836486    
## student                           1   0.11  0.1100   0.784 0.376531    
## gender                            1   0.18  0.1800   1.282 0.258186    
## format:conference                 1   0.05  0.0498   0.355 0.551780    
## format:student                    1   0.21  0.2150   1.532 0.216638    
## conference:student                1   0.00  0.0004   0.003 0.957688    
## format:gender                     1   0.03  0.0261   0.186 0.666379    
## conference:gender                 1   0.03  0.0252   0.179 0.672053    
## student:gender                    1   0.07  0.0656   0.468 0.494537    
## format:conference:student         1   0.10  0.0987   0.704 0.402147    
## format:conference:gender          1   0.00  0.0046   0.032 0.857041    
## format:student:gender             1   0.05  0.0546   0.389 0.533309    
## conference:student:gender         1   0.02  0.0195   0.139 0.709404    
## format:conference:student:gender  1   0.03  0.0290   0.206 0.649913    
## Residuals                       368  51.65  0.1403                     
## ---
## 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",])
summary(m0)
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## dalechall_score_norm ~ 1 + (format + student + gender + review)^2 +
##     time + (1 + format + gender | conference)
##    Data: readScores[readScores$conference != "E8", ]
##
## REML criterion at convergence: 699.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5716 -0.6737  0.0250  0.6070  3.0111
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.  Corr
##  conference (Intercept) 0.000e+00 0.000e+00
##             format1     1.743e-16 1.320e-08   NaN
##             gender1     1.374e-16 1.172e-08   NaN -1.00
##  Residual               1.400e-01 3.742e-01
## Number of obs: 753, groups:  conference, 4
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)             6.108027   0.049189 741.000000 124.174  < 2e-16
## format1                 0.163237   0.051187 741.000000   3.189  0.00149
## student1                0.011899   0.047173 741.000000   0.252  0.80093
## gender1                -0.046606   0.046148 741.000000  -1.010  0.31286
## reviewSingle           -0.026703   0.066005 741.000000  -0.405  0.68592
## time                    0.007782   0.027609 741.000000   0.282  0.77812
## format1:student1        0.054107   0.070616 741.000000   0.766  0.44380
## format1:gender1         0.023343   0.068188 741.000000   0.342  0.73220
## format1:reviewSingle   -0.003249   0.062427 741.000000  -0.052  0.95850
## student1:gender1       -0.040472   0.062358 741.000000  -0.649  0.51652
## student1:reviewSingle  -0.012838   0.060627 741.000000  -0.212  0.83235
## gender1:reviewSingle    0.044642   0.058281 741.000000   0.766  0.44393
##
## (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
```

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:

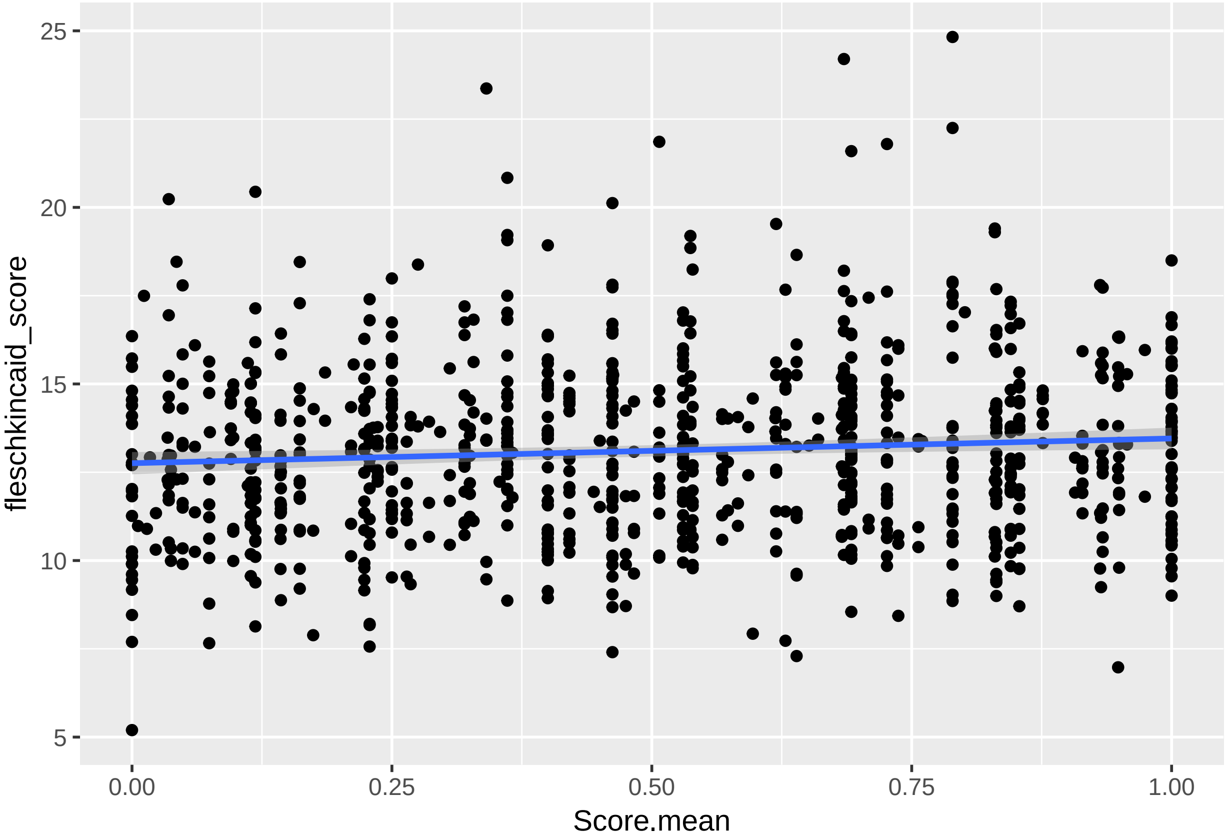
```
cor.test(readScores$Score.mean, readScores$fleschkincaid_score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$fleschkincaid_score
## t = 2.5828, df = 900, p-value = 0.009956
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02061699 0.15021152
## sample estimates:
##        cor
## 0.08577706
```

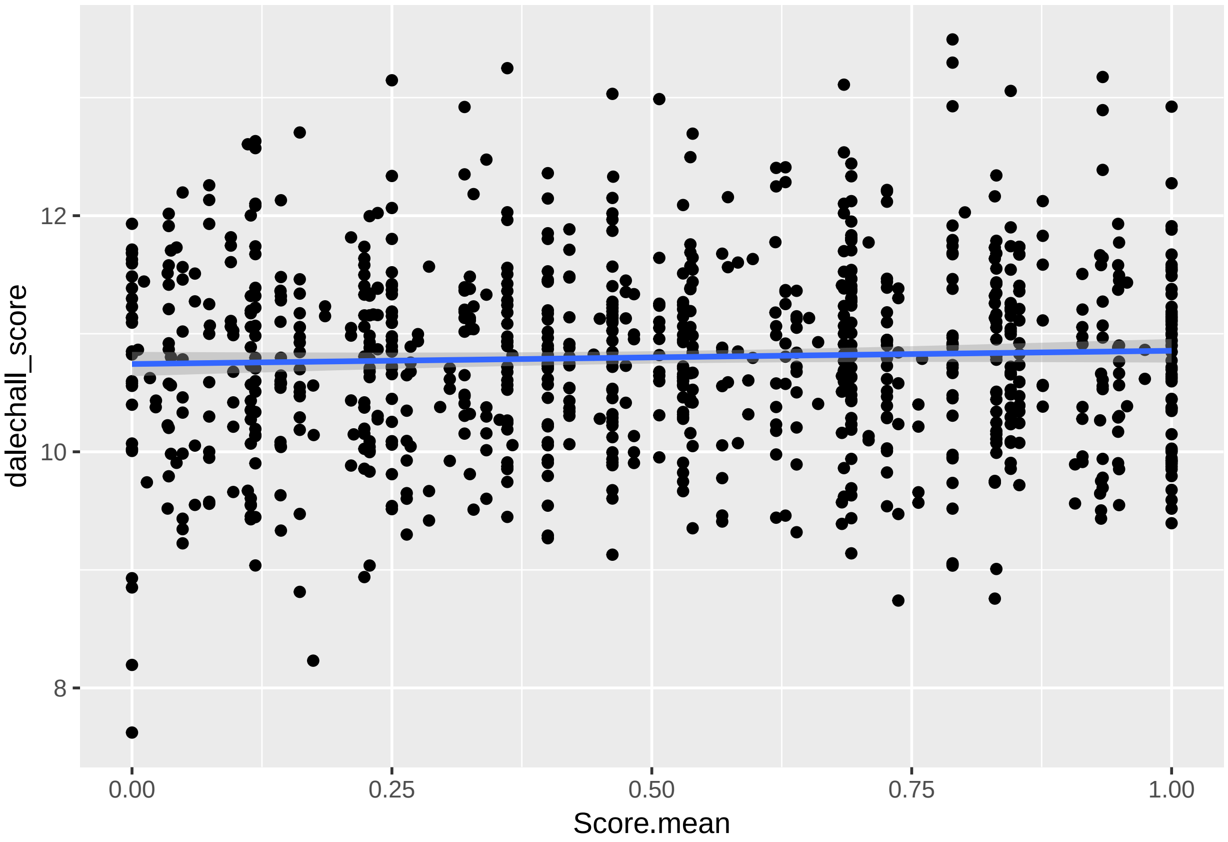
```
cor.test(readScores$Score.mean, readScores$dalechall_score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$dalechall_score
## t = 1.2699, df = 900, p-value = 0.2044
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02304636  0.10727214
## sample estimates:
##        cor
## 0.04229277
```

```
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)
m1 = update(m0,~.+fleschkincaid_score_scaled)
m2 = update(m1,~.+fleschkincaid_score_scaled:gender)
anova(m0,m1,m2)
```

```
## Data: readScores
## Models:
## object: Score.mean.norm ~ 1 + format + student + gender + (1 | conference)
## ..1: Score.mean.norm ~ format + student + gender + (1 | conference) +
## ..1:     fleschkincaid_score_scaled
## ..2: Score.mean.norm ~ format + student + gender + (1 | conference) +
## ..2:     fleschkincaid_score_scaled + gender:fleschkincaid_score_scaled
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## object  6 2126.5 2154.2 -1057.2   2114.5
## ..1     7 2127.1 2159.4 -1056.5   2113.1 1.4443      1     0.2294
## ..2     8 2129.0 2166.0 -1056.5   2113.0 0.0182      1     0.8926
```

```
summary(m2)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## 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
##   2129.0   2166.0  -1056.5   2113.0      745
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9234 -0.9003 -0.0175  0.8968  1.9443
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  conference (Intercept) 0.0000   0.0000
##  Residual               0.9688   0.9843
## Number of obs: 753, groups:  conference, 4
##
## Fixed effects:
##                                     Estimate Std. Error t value
## (Intercept)                        -0.02976    0.04538  -0.656
## format1                             0.24568    0.08270   2.971
## student1                           -0.01985    0.07807  -0.254
## gender1                             0.12686    0.07535   1.684
## fleschkincaid_score_scaled          0.04840    0.04147   1.167
## gender1:fleschkincaid_score_scaled  0.01113    0.08244   0.135
##
## Correlation of Fixed Effects:
##             (Intr) formt1 stdnt1 gendr1 flsc__
## format1     -0.468
## student1    -0.360  0.091
## gender1      0.273 -0.116  0.024
## flschkncd__  0.032 -0.136 -0.004 -0.017
## gndr1:fls__ -0.072  0.061  0.055 -0.041  0.370
```

Dale-Chall scores:

```
m0 = lmer(Score.mean.norm~ 1 +
            format + student + gender +
           (1 | conference),
       data = readScores,
       REML = F)
m1 = update(m0,~.+dalechall_score_scaled)
m2 = update(m1,~.+dalechall_score_scaled:gender)
anova(m0,m1,m2)
```

```
## Data: readScores
## Models:
## object: Score.mean.norm ~ 1 + format + student + gender + (1 | conference)
## ..1: Score.mean.norm ~ format + student + gender + (1 | conference) +
## ..1:     dalechall_score_scaled
## ..2: Score.mean.norm ~ format + student + gender + (1 | conference) +
## ..2:     dalechall_score_scaled + gender:dalechall_score_scaled
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## object  6 2126.5 2154.2 -1057.2   2114.5
## ..1     7 2128.5 2160.8 -1057.2   2114.5 0.0117      1     0.9140
## ..2     8 2130.5 2167.5 -1057.2   2114.5 0.0158      1     0.8999
```

```r
summary(m2)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Score.mean.norm ~ format + student + gender + (1 | conference) +
##     dalechall_score_scaled + gender:dalechall_score_scaled
##    Data: readScores
##
##      AIC      BIC   logLik deviance df.resid
##   2130.5   2167.5  -1057.2   2114.5      745
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.89072 -0.90090 -0.03418  0.91063  1.89314
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  conference (Intercept) 0.0000   0.0000
##  Residual               0.9706   0.9852
## Number of obs: 753, groups:  conference, 4
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                    -0.032909   0.045329  -0.726
## format1                         0.263319   0.083196   3.165
## student1                       -0.017948   0.078066  -0.230
## gender1                         0.127700   0.075471   1.692
## dalechall_score_scaled         -0.004810   0.038251  -0.126
## gender1:dalechall_score_scaled -0.009455   0.075132  -0.126
##
## Correlation of Fixed Effects:
##             (Intr) formt1 stdnt1 gendr1 dlch__
## format1     -0.466
## student1    -0.359  0.090
## gender1      0.274 -0.121  0.024
## dlchll_scr_  0.059 -0.194 -0.027  0.021
## gndr1:dlc__ -0.028  0.041  0.025 -0.046  0.151
```

No interactions.

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:

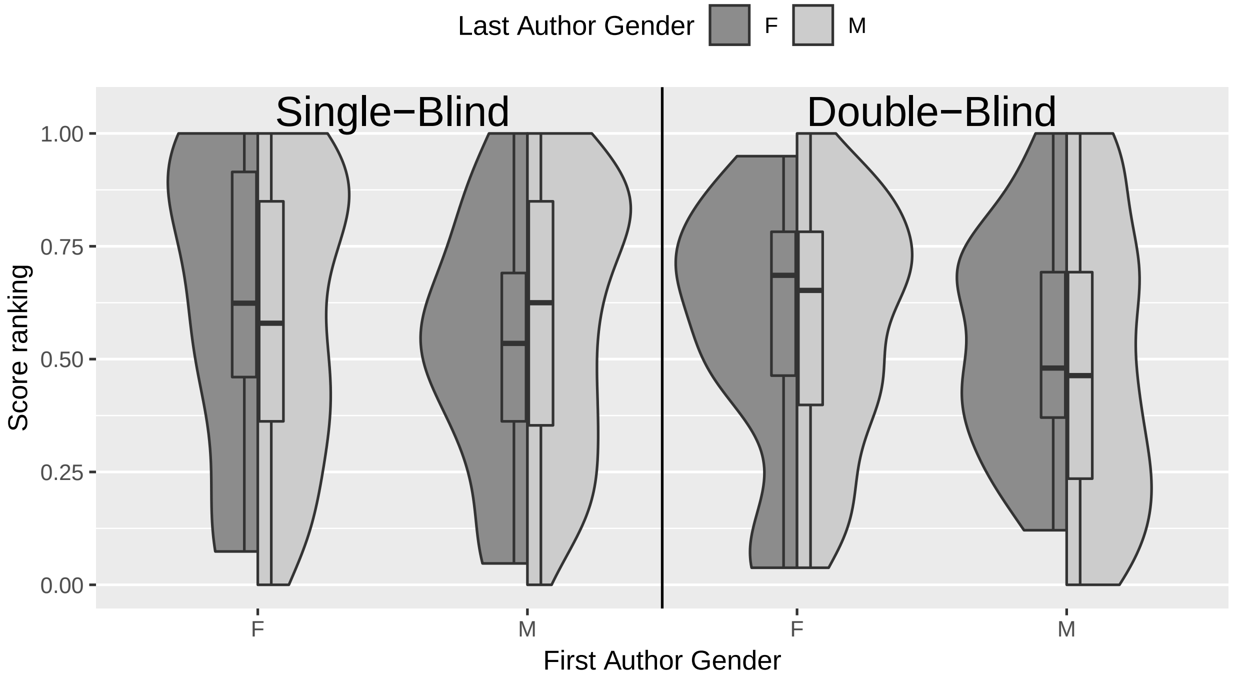


Figure 1: Distributions of review scores by first and last author gender.

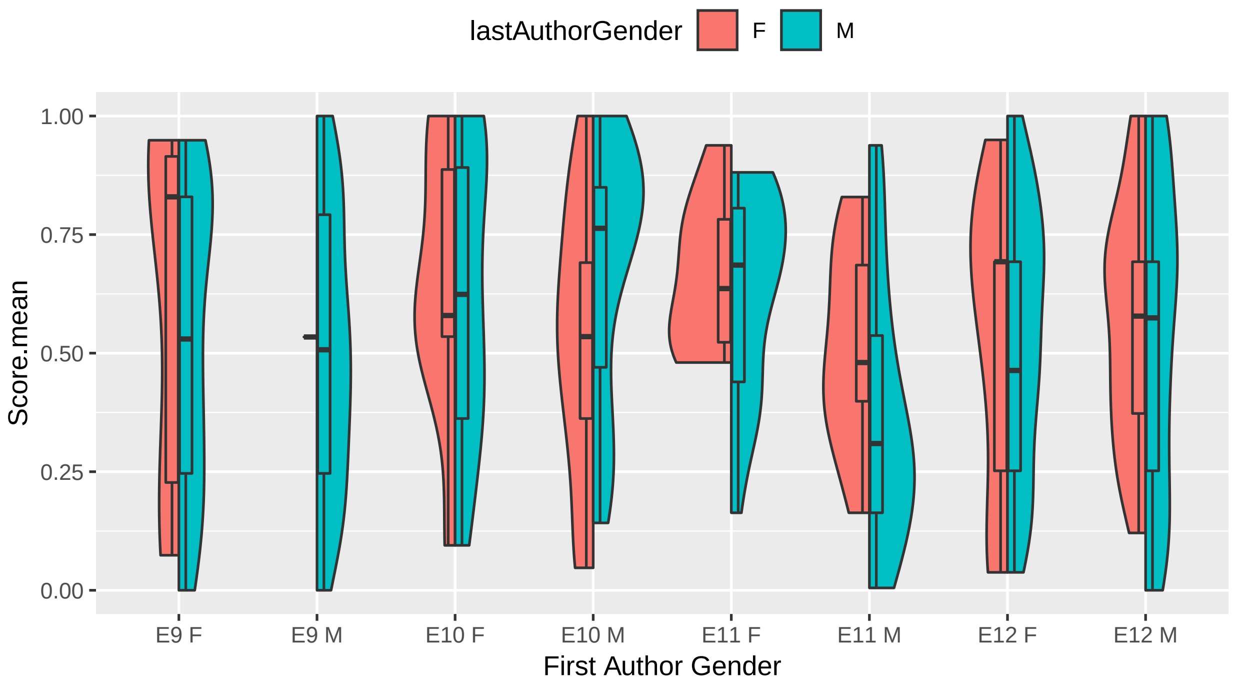


Figure 2: Distributions of review scores by first and last author gender.

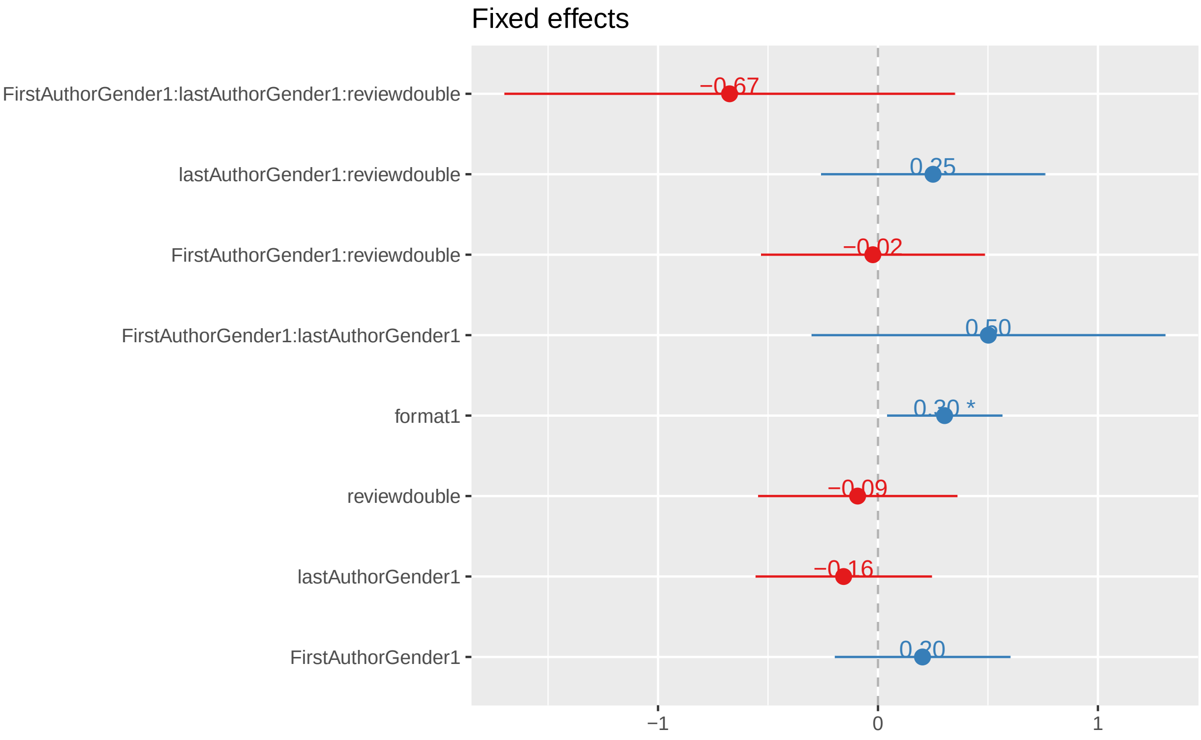


Figure 3: Coefficients and confidence intervals for effects predicting review ranks.