Readability of submissions to EvoLang

Contents

Introdution

Study 1 found that the effect of double-blind peer-review at EvoLang 11 did not persist significantly at EvoLang 12. The results of EvoLang 11 may have been an anomaly, or caused by some other factor that differs between the conferences (proportion of genders, location, different authors, etc.). Another possibility is that the advantage for female authors in EvoLang 11 occurred because they had better writing (as suggested by Hengel, 2017). Male authors may have changed their strategy after having experienced double-blind review (or they may have read Roberts & Verhoef, 2016; though see Handley et al., 2015b) by investing more effort into writing their submissions for EvoLang 12. The readability study tests this by measuring the readability of submissions, assessing whether the readability of male and female authors differs significantly between single-blind and double-blind conferences.

Data

Text from submissions was extracted automatically from pdf or Microsoft Word formats using the command line programs `textutil` and `pdftotext`. The texts were cleaned to remove various features (author names, affiliations, titles, bibliography, acknowledgements, reference manager artefacts, decimal characters, references to figures, figure captions, tables and linguistic examples). Readability scores for submissions were calculated using the code supplied in Hengel (2017). 902 submissions from 5 conferences (97%) could be analysed. Based on their relative independence in our sample, two measures were analysed: the Flesch-Kincaid grade level and the Dale-Chall readability formula. The Flesch-Kincaid score estimates the US school grade level required to understand the text. Dale-Chall readability corresponds with US grade level less straightforwardly, with any score above 10 requiring a university-level vocabulary for understanding. In both cases, a higher score indicates that the text is more complicated.

Results

Table 1 shows the mean Flesch-Kincaid scores by conference and gender. The mean Flesch-Kincaid score was 13.2 (sd = 2.53, see figure 1) and the mean Dale-Chall score was 10.8 (sd = 0.8, correlation between them r = 0.63). We note that there is relatively little variation in these samples.

Data for conferences 9-12 were analysed with a mixed effects model with random intercepts for each conference and random slopes for submission type, student status and gender (see SI). A continuous fixed effect representing the year that the conference was held was added to test whether readability was changing

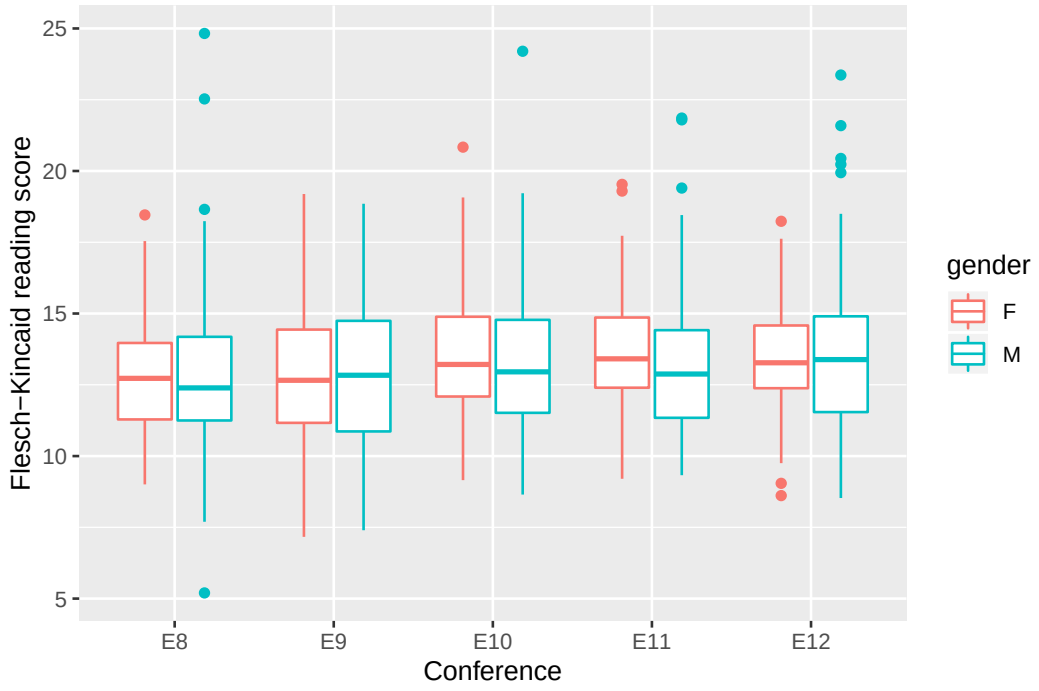


Figure 1: Flesch-Kincaid reading scores by conference and gender.

over time. There was no significant effect of author gender. The only significant effect was that abstracts had higher reading scores than full papers.

An ANOVA did not find evidence that the difference in readability between men and women in E11 (males were 0.49 Flesch-Kincaid points below females) was significantly bigger than the corresponding difference in E12 (males were 0.12 Flesch-Kincaid points above females; Flesch-Kincaid: $F(1) = 0.56$, $p = 0.45$; Dale-Chall: $F(1) = 0.11$, $p = 0.73$).

There was a very weak correlation between the readability scores and the reviewer scores and this was non-significant for the Dale-Chall score (Flesch-Kincaid $r = 0.08$, $p = 0.01$; Dale-Chall $r = 0.05$, $p = 0.10$).

```
knitr::include_graphics("../results/FleschKincaidReadingScores2.pdf")
```

E8	E9	E10	E11	E12	
Female	12.92 (2.16)	12.69 (2.56)	13.45 (2.36)	13.61 (2.2)	13.39 (1.85)
Male	12.8 (2.85)	12.9 (2.43)	13.31 (2.43)	13.12 (2.31)	13.5 (2.75)

Table 1: Mean Flesch-Kincaid reading scores (and standard deviations in parentheses) by conference and gender.

Conclusion

The readability analysis did not show strong evidence for a difference in readability by gender or review type. Hengel (2017) only found a 1-6% difference in readability scores, which might not reach significance in our smaller sample.

References

Handley, I. M., Brown, E. R., Moss-Racusin, C. A., & Smith, J. L. (2015b). Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. Proceedings of the National Academy of Sciences, 112(43), 13201-13206.

Hengel, E. (2017). Publishing while Female. Are women held to higher standards? Evidence from peer review. Cambridge Working Papers in Economics 1753, Faculty of Economics, University of Cambridge. https://ideas.repec.org/p/cam/camdae/1753.html

Statistical 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)
```

Load data

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            -0.73                0.63             0.62
##                     smog_score dalechall_score
## flesch_score             -0.93           -0.73
## fleschkincaid_score       0.96            0.63
## 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   52 129
##   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 84  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.

Note that there is one outlier paper with a Flesch-Kincaid score of 34. The text was checked, and there were no transcription errors. This paper has more than 350 words in 5 sentences, more than three times the average words per sentence. There are also some very short abstracts, mainly from EvoLang 8 where the format was less well established.

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

```
## [1] 13.18095
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  readScores$fleschkincaid_score and readScores$dalechall_score
## t = 24.387, df = 899, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5899851 0.6687359
## sample estimates:
##       cor
## 0.6309832
```

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

```
## [1] -0.4857825
```

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

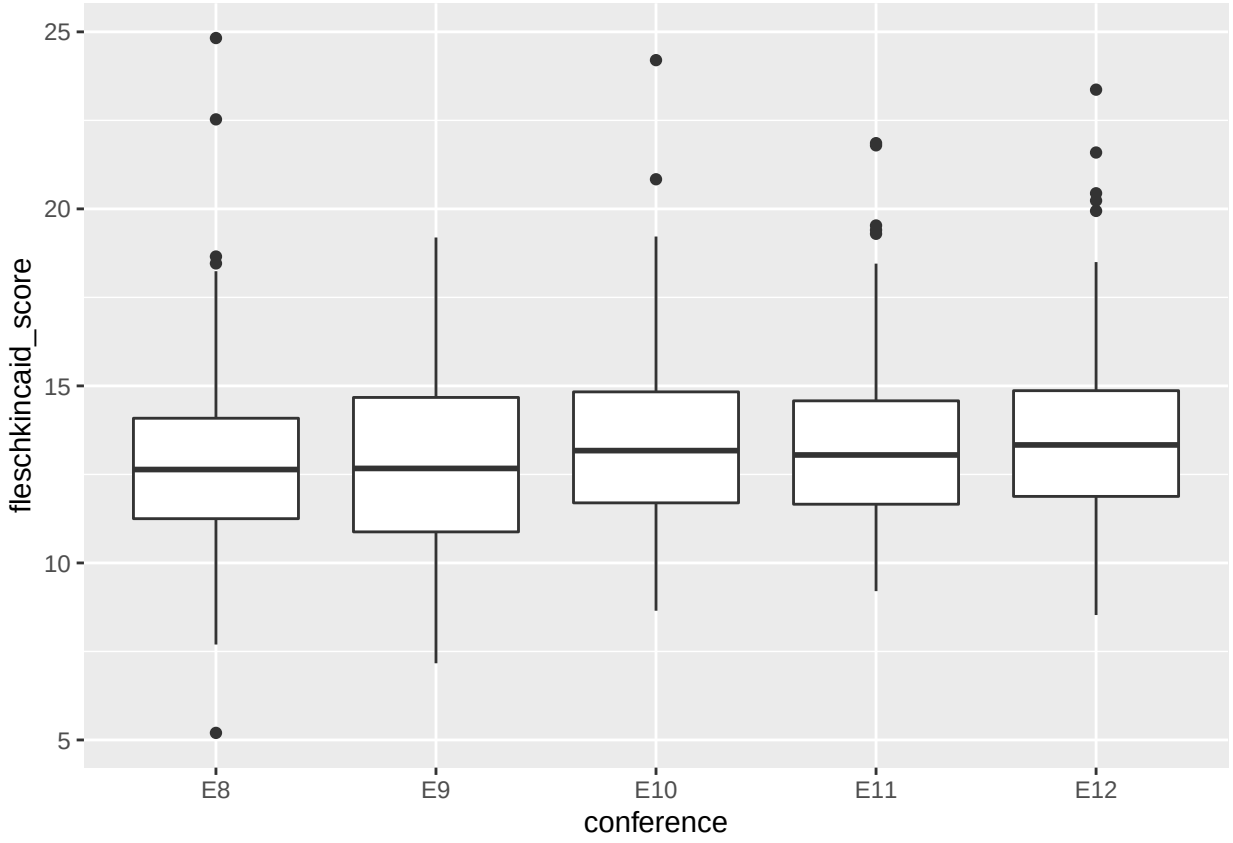
```
## [1] 0.1066118
```

```
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))
```

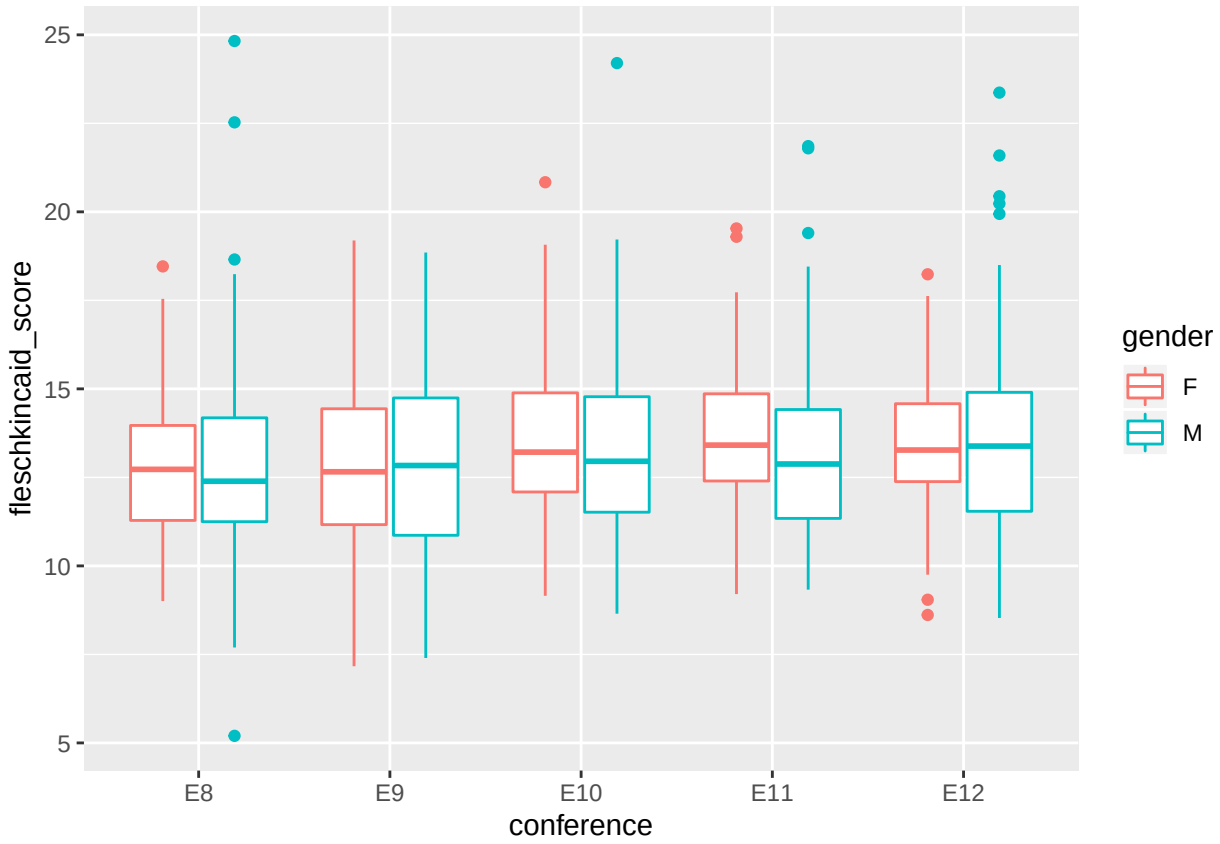
```
msdFK = matrix(paste0(round(meanFK,2)," (",round(sdFK,2),")"),nrow=2)
colnames(msdFK) = sort(unique(readScores$conference))
rownames(msdFK) = c("Female","Male")
write.csv(msdFK,"../results/MeanFleschKincaidScores_by_conf_by_gender.csv")
```

Various Plots:

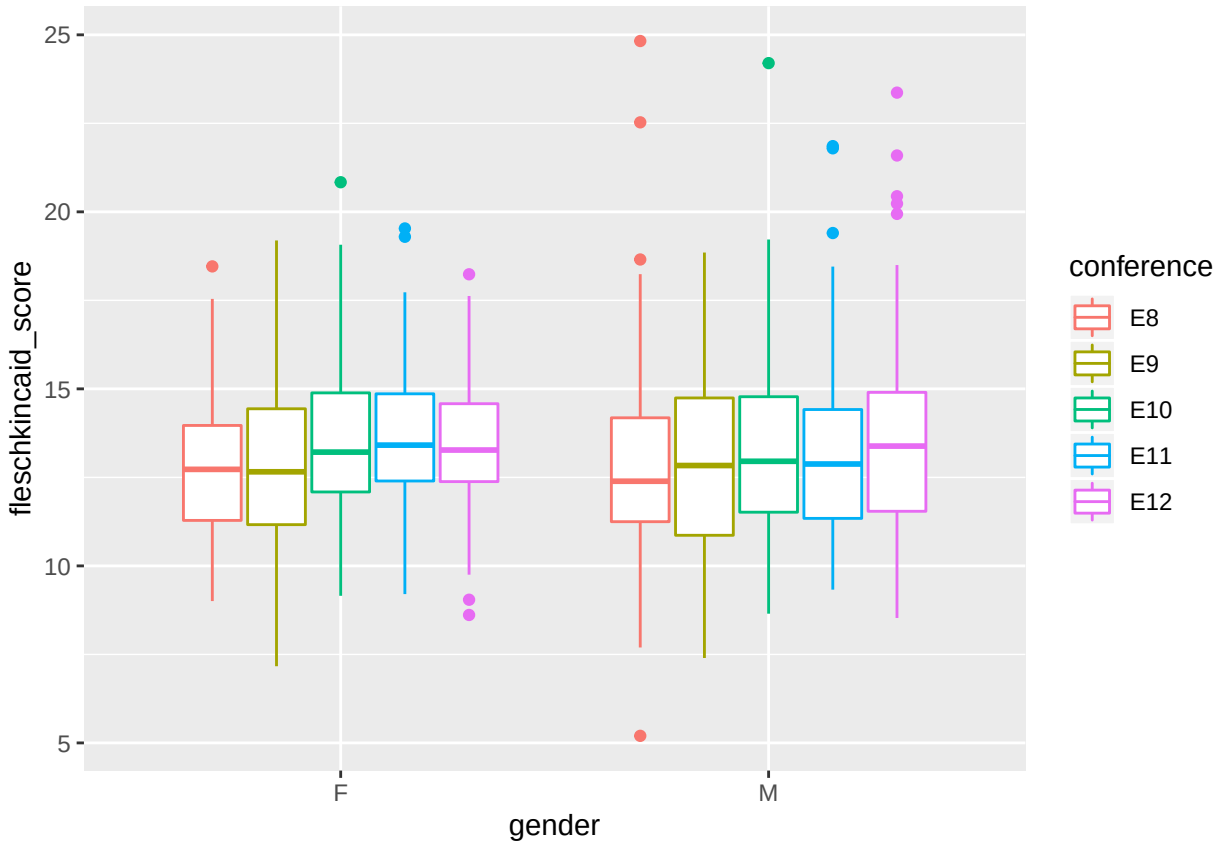
```
readScores$gender2 = "Female"
readScores$gender2[readScores$gender=="M"] = "Male"

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



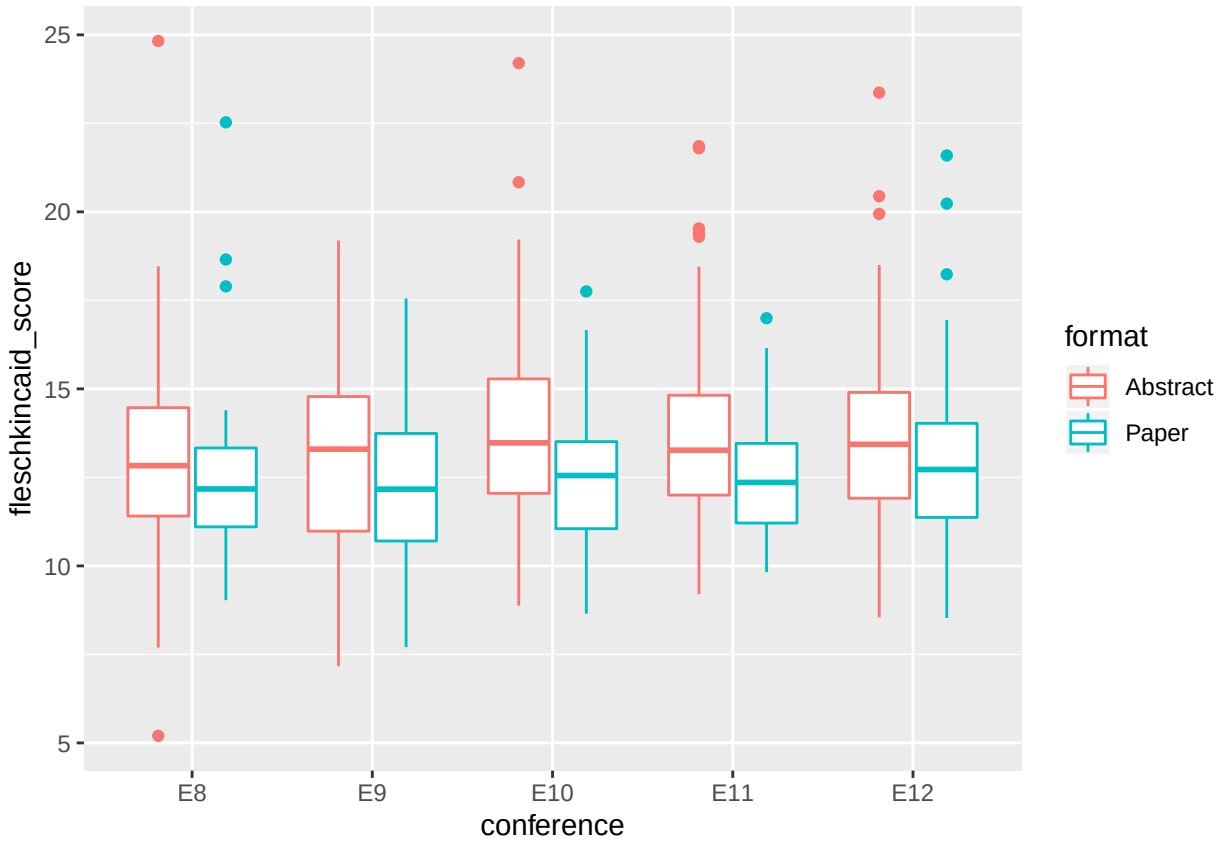

```
ggplot(readScores, aes(y=fleschkincaid_score,x=conference,colour=gender)) + geom_boxplot()
```



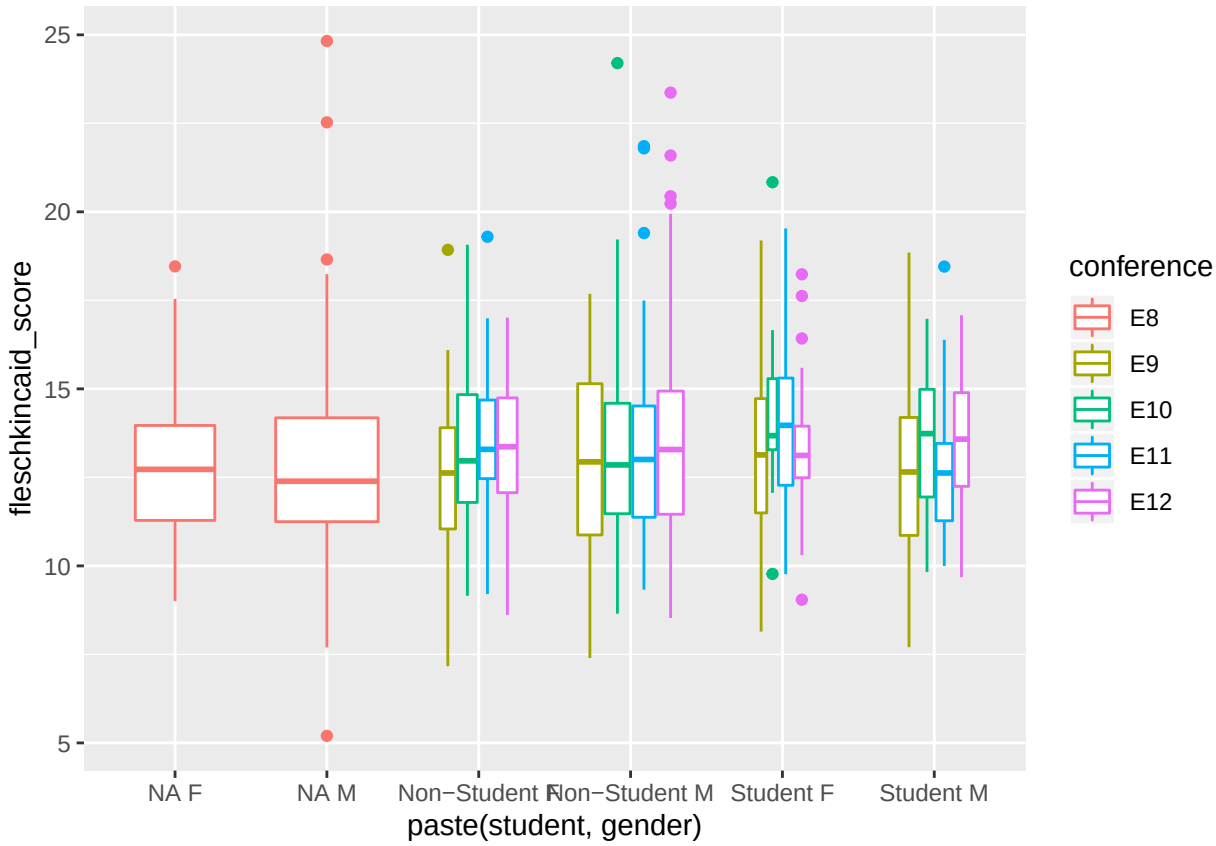

```
ggplot(readScores, aes(y=fleschkincaid_score,x=gender,colour=conference)) + geom_boxplot()
```




```
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(
```



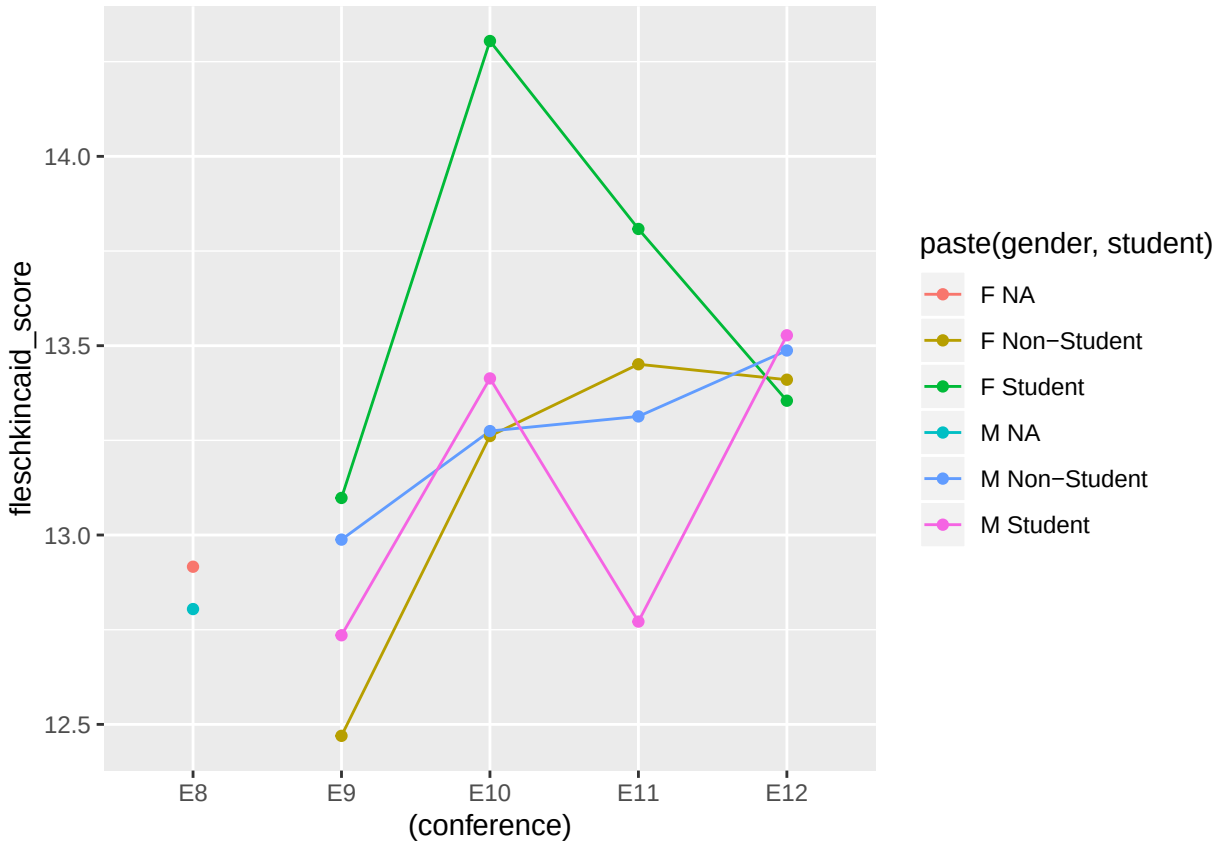

```
fkrs = ggplot(readScores, aes(y=fleschkincaid_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Flesch-Kincaid reading score", x="Gender")

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

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

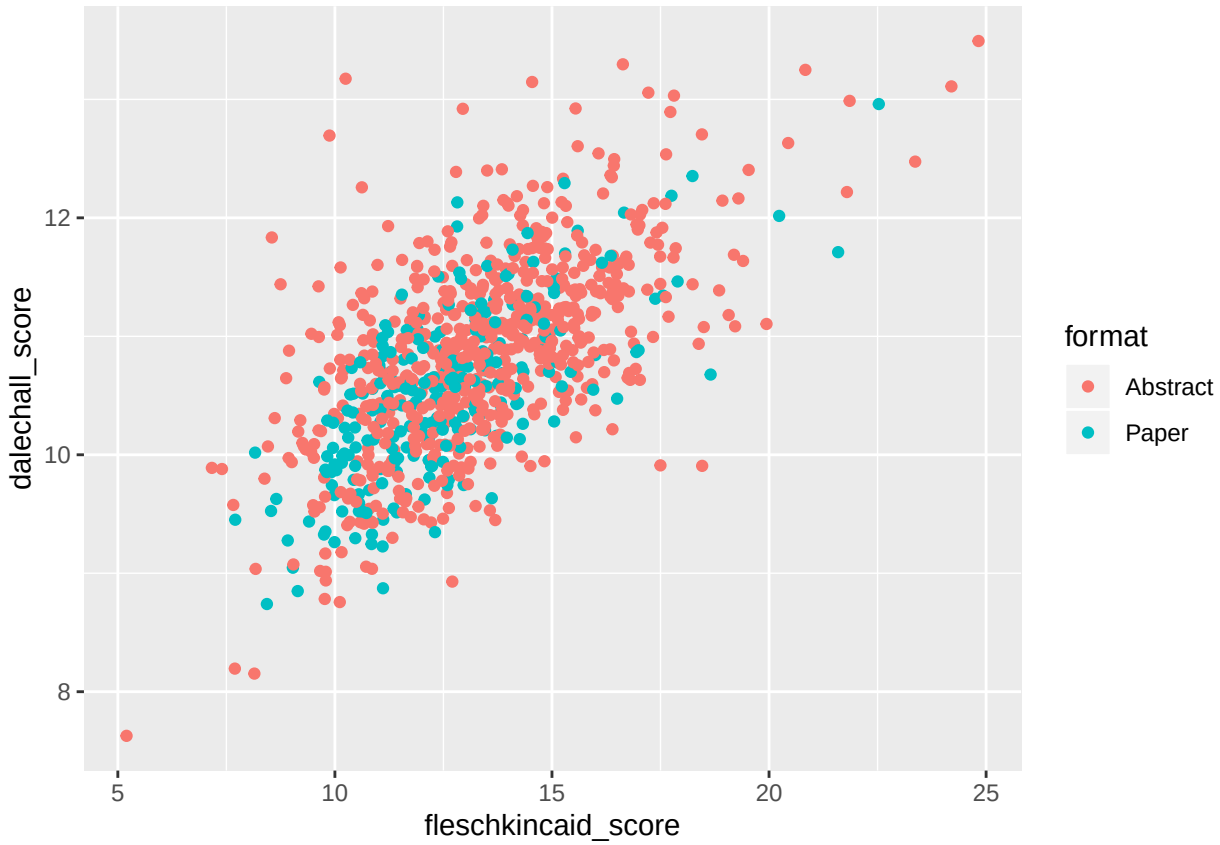
```
x = readScores %>% group_by(conference,gender,student) %>%
  summarise(dalechall_score=mean(dalechall_score),
            fleschkincaid_score=mean(fleschkincaid_score))
```

```
## 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()
```

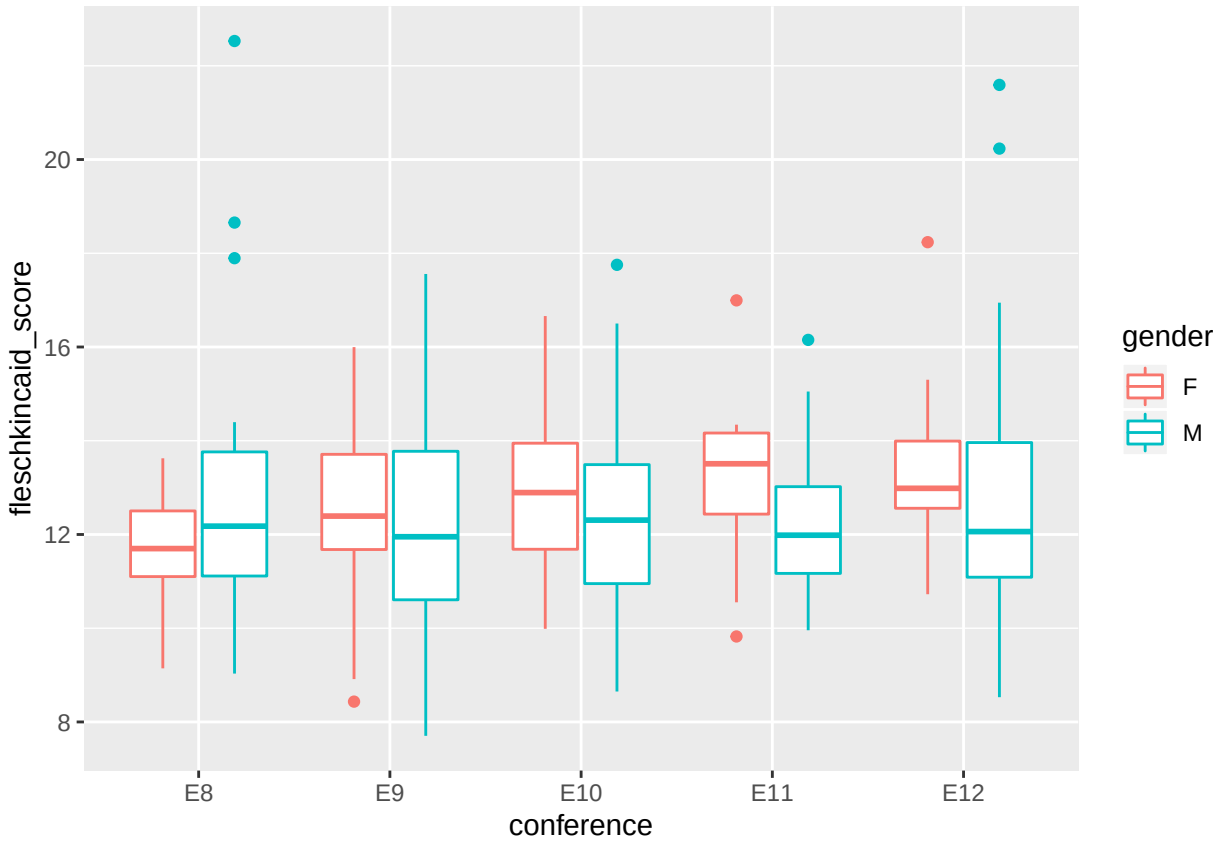



```
fkrs2= ggplot(readScores, aes(y=fleschkincaid_score,
                              x=conference,colour=gender)) +
  geom_boxplot() +
  labs(y="Flesch-Kincaid reading score", x="Conference")

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

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

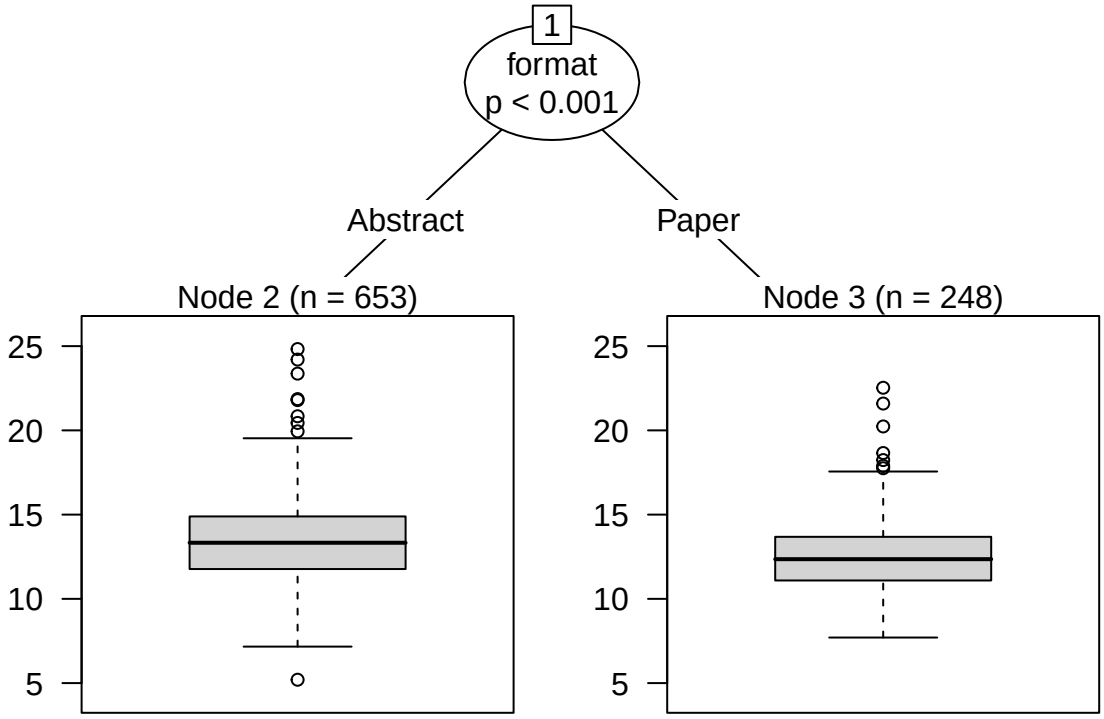
Flesch-Kincaid score for full papers only:

```
ggplot(readScores[readScores$format=="Paper",],
       aes(x=conference,y=fleschkincaid_score,colour=gender)) + geom_boxplot()
```


fleschkincaid_score
20
16
12
8
E8
E9
E10
E11
E12
conference
gender
F
M

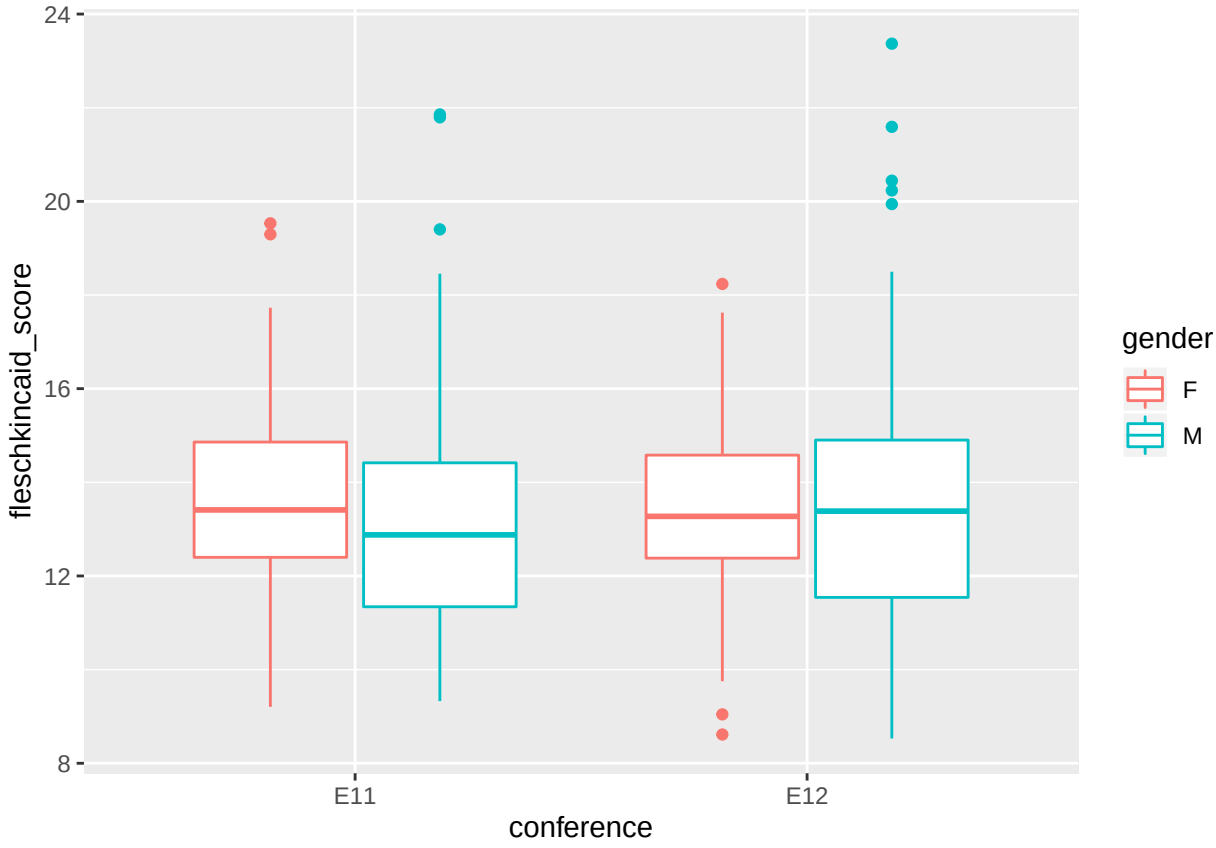
Decision tree

```
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  0.486  0.4863  10.025 0.00167 **
## conference                        1  0.007  0.0068   0.141 0.70763
## student                           1  0.015  0.0151   0.311 0.57736
## gender                            1  0.017  0.0175   0.360 0.54895
## format:conference                 1  0.046  0.0458   0.945 0.33175
## format:student                    1  0.016  0.0158   0.325 0.56902
## conference:student                1  0.019  0.0195   0.402 0.52671
## format:gender                     1  0.099  0.0990   2.040 0.15401
## conference:gender                 1  0.027  0.0275   0.567 0.45212
## student:gender                    1  0.035  0.0353   0.728 0.39420
## format:conference:student         1  0.042  0.0419   0.864 0.35336
## format:conference:gender          1  0.000  0.0000   0.000 0.99246
## format:student:gender             1  0.000  0.0004   0.008 0.92672
## conference:student:gender         1  0.046  0.0456   0.940 0.33285
## format:conference:student:gender  1  0.125  0.1251   2.579 0.10912
## Residuals                       368 17.850  0.0485
## ---
```

```
## 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: 2051.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4914 -0.7031 -0.0655  0.6052  4.4346
##
## Random effects:
##  Groups     Name        Variance  Std.Dev. Corr
##  conference (Intercept) 0.0035040 0.05919
##             format1     0.0070358 0.08388   1.00
##             student1    0.0000354 0.00595  -1.00 -1.00
##             gender1     0.0066688 0.08166   1.00  1.00 -1.00
##  Residual               0.8691361 0.93227
## Number of obs: 752, groups:  conference, 4
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)            -0.197345   0.149055   3.793772  -1.324   0.2596
## format1                 0.295531   0.140709   7.181899   2.100   0.0728 .
## student1                0.059603   0.117693 396.680638   0.506   0.6128
## gender1                 0.155111   0.128742   6.093458   1.205   0.2730
## reviewSingle            0.184804   0.201695   3.687153   0.916   0.4155
## time                    0.122890   0.085386   3.624831   1.439   0.2305
## format1:student1       -0.012731   0.176303 732.628341  -0.072   0.9425
## format1:gender1        -0.271648   0.170196 736.217912  -1.596   0.1109
## format1:reviewSingle    0.008015   0.176907   4.496672   0.045   0.9658
## student1:gender1       -0.274790   0.155549 727.633848  -1.767   0.0777 .
## student1:reviewSingle  -0.155118   0.151572 292.589510  -1.023   0.3070
## gender1:reviewSingle   -0.067521   0.166779   4.279872  -0.405   0.7050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
```

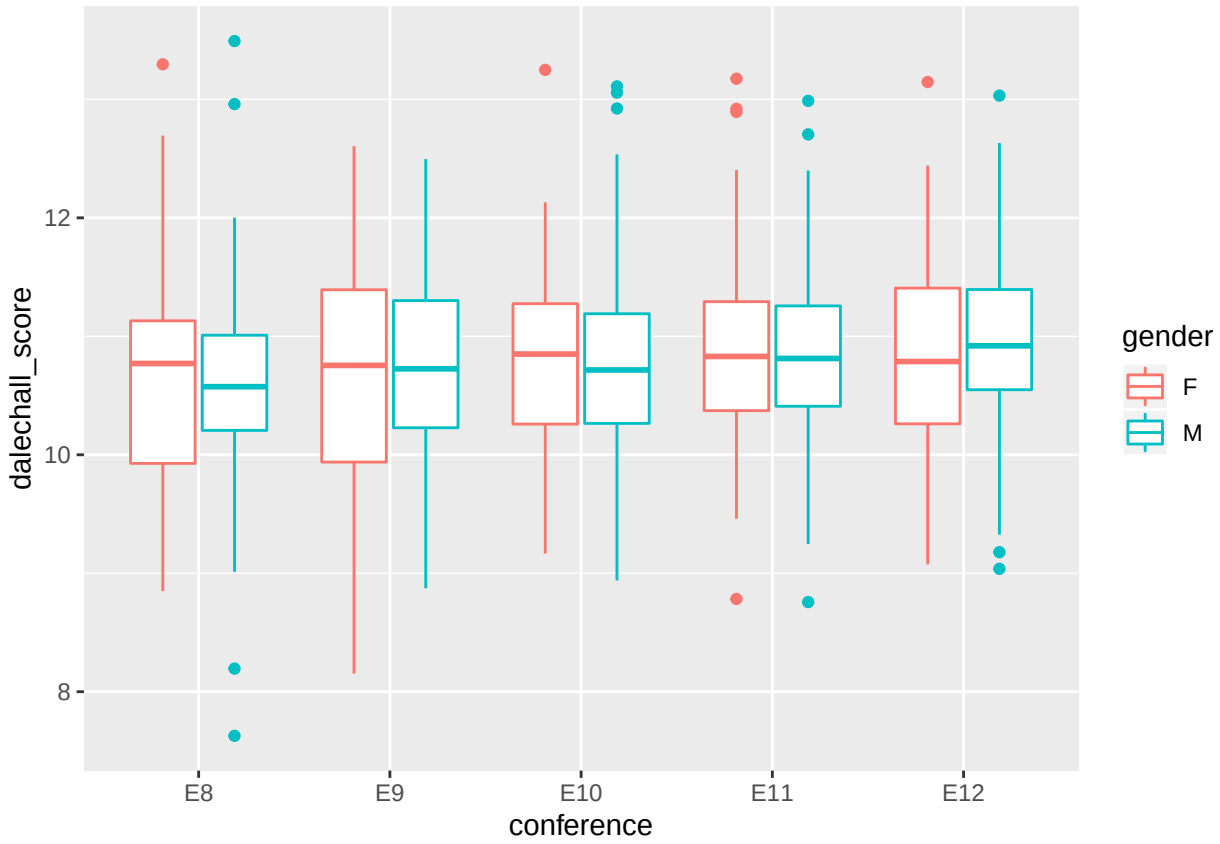
```
##             (Intr) formt1 stdnt1 gendr1 rvwSng time   frmt1:s1 frmt1:g1
## format1     -0.101
## student1    -0.114  0.207
## gender1      0.264  0.005  0.033
## reviewSingl -0.917  0.041  0.042 -0.175
## time        -0.852 -0.034 -0.071 -0.037  0.850
## frmt1:stdn1  0.074 -0.301 -0.481 -0.003  0.003  0.045
## frmt1:gndr1 -0.150  0.235 -0.018 -0.422  0.077  0.061  0.030
## frmt1:rvwSn  0.044 -0.680 -0.058 -0.083 -0.023  0.016  0.016    0.002
## stdnt1:gnd1 -0.012  0.014  0.126 -0.200 -0.034  0.022 -0.077    0.058
## stdnt1:rvwS  0.122 -0.088 -0.633 -0.058 -0.186 -0.012  0.120    0.001
## gndr1:rvwSn -0.154 -0.052 -0.041 -0.638  0.248 -0.009 -0.002    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  0.053  0.125
## gndr1:rvwSn  0.150 -0.073 -0.016
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

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

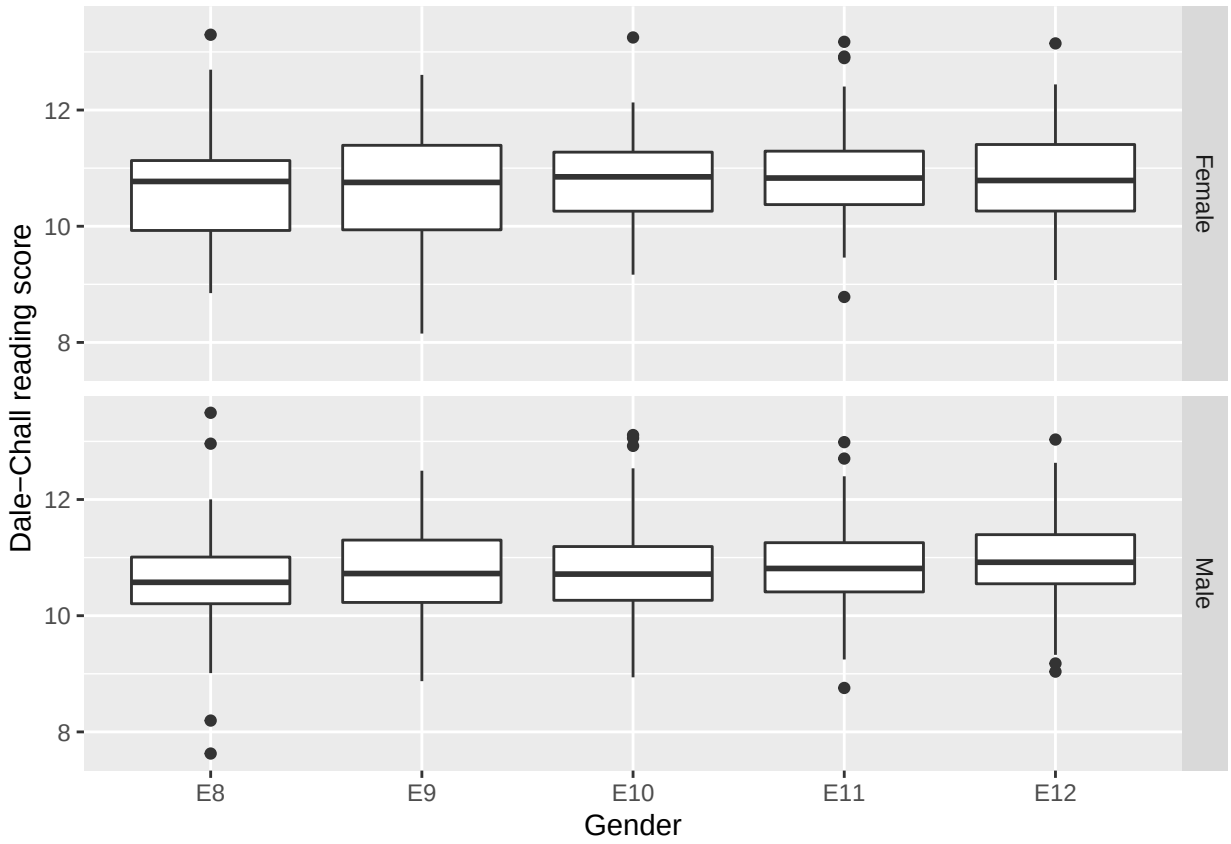
Dale-Chall scale

Plots

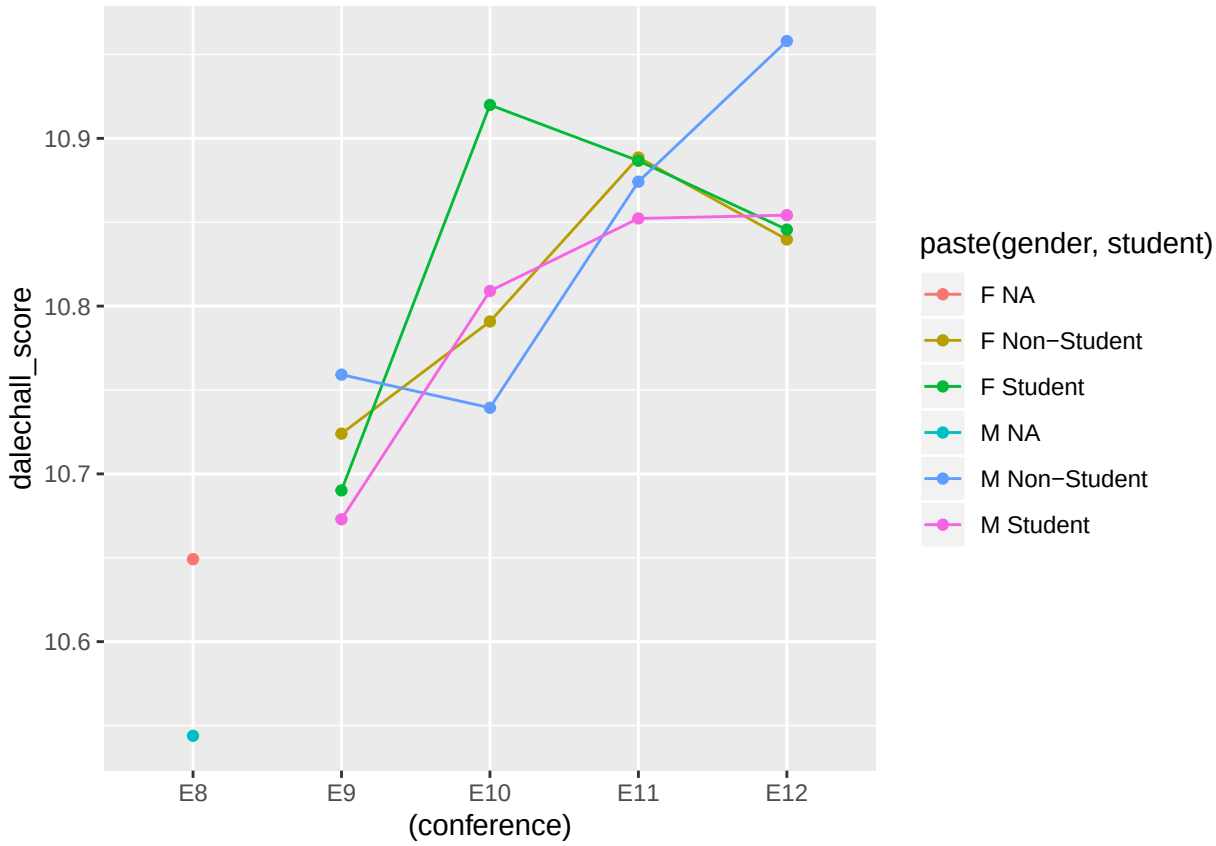
```
ggplot(readScores, aes(y=dalechall_score,x=conference,colour=gender)) + geom_boxplot()
```



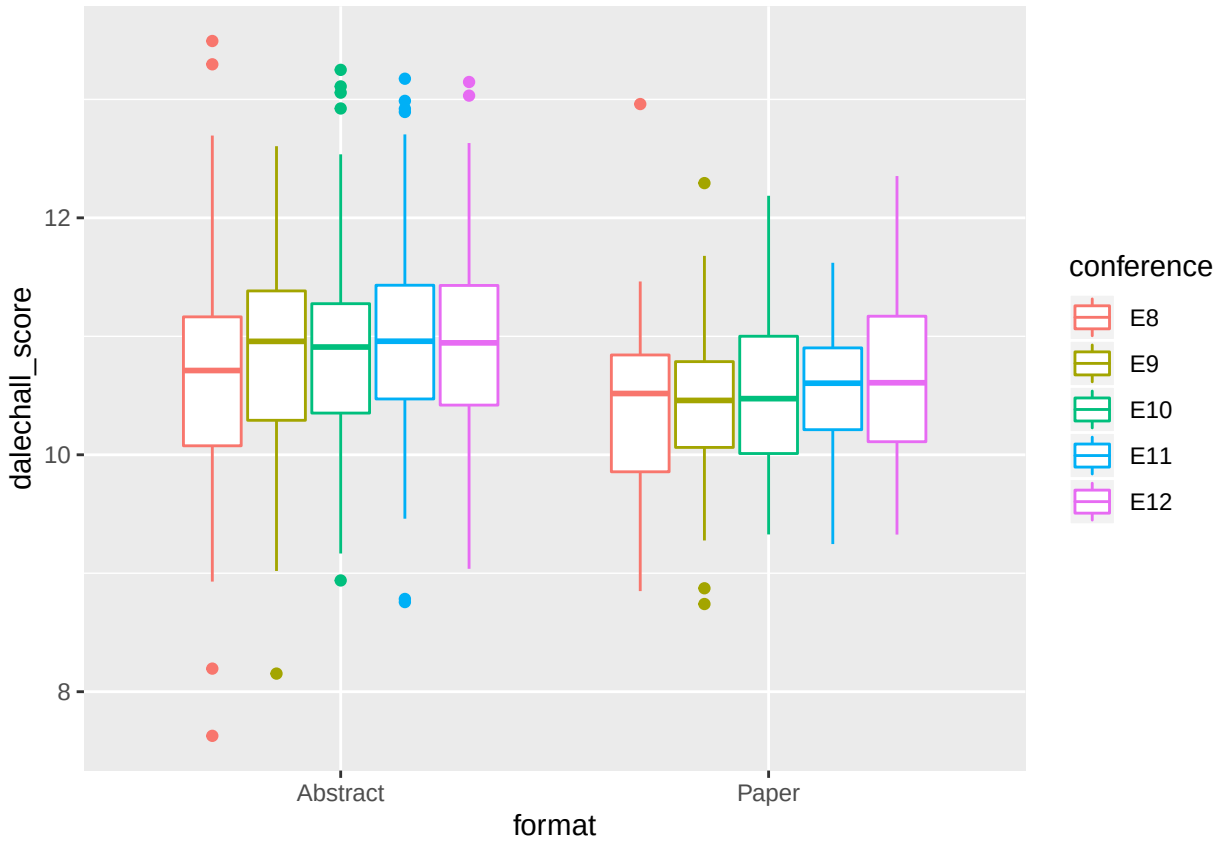

```
ggplot(readScores, aes(y=dalechall_score,x=conference)) +
  geom_boxplot() + facet_grid("gender2") +
  labs(y="Dale-Chall reading score", x="Gender")
```




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

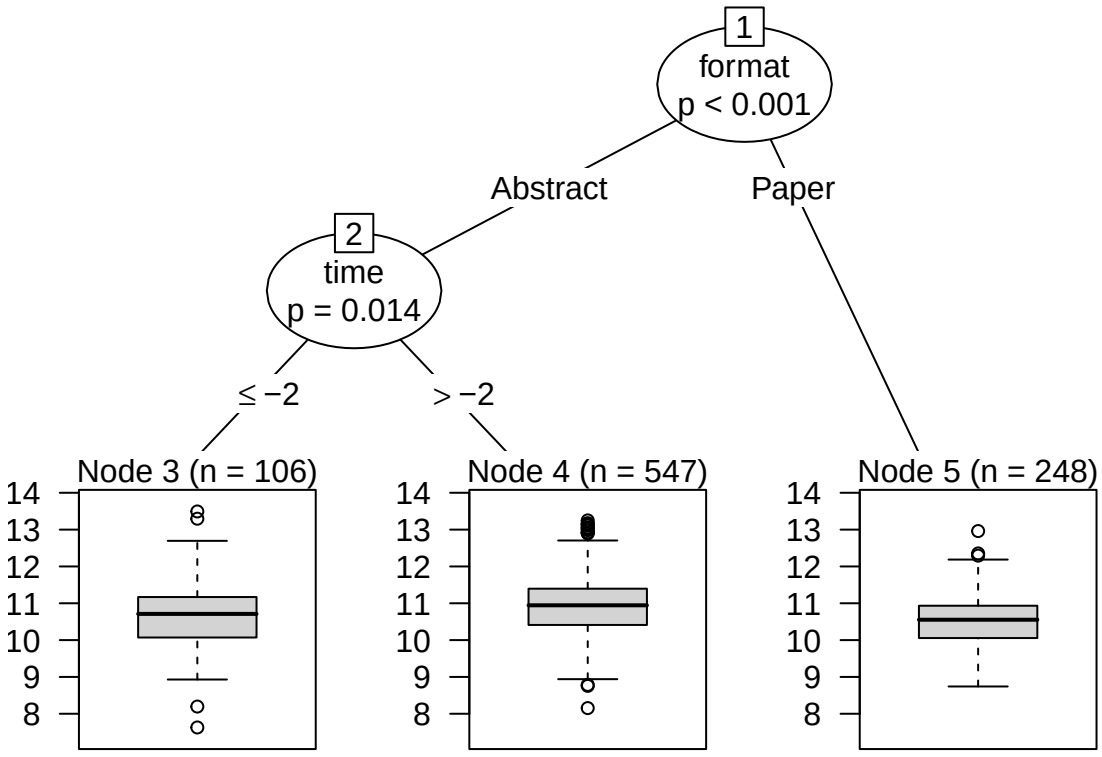



```
ggplot(readScores, aes(y=dalechall_score,x=format,colour=conference)) + geom_boxplot()
```


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

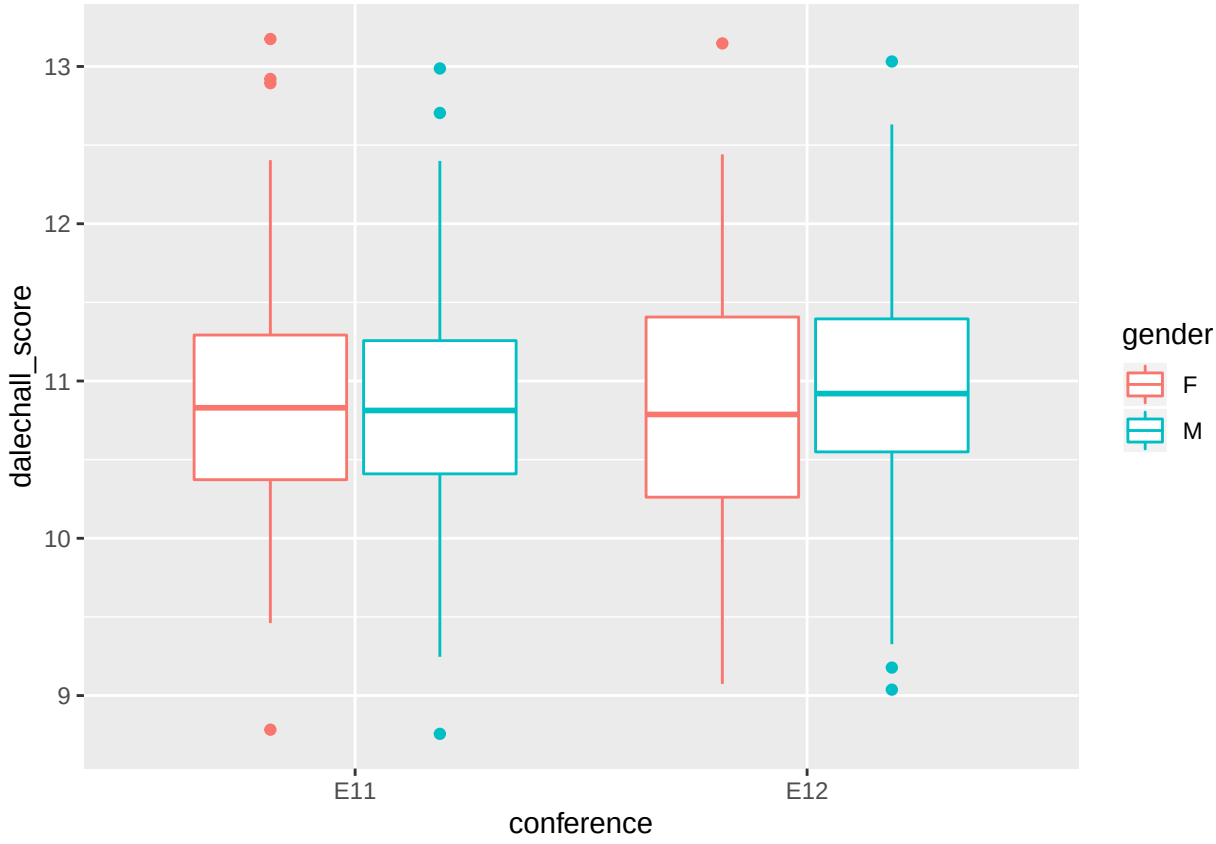
Decision tree:

```
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   2.13  2.1328  14.999 0.000127 ***
## conference                        1   0.00  0.0008   0.005 0.941248
## student                           1   0.14  0.1352   0.951 0.330179
## gender                            1   0.15  0.1483   1.043 0.307820
## format:conference                 1   0.08  0.0820   0.577 0.448062
## format:student                    1   0.19  0.1918   1.349 0.246267
## conference:student                1   0.00  0.0000   0.000 0.987339
## format:gender                     1   0.07  0.0715   0.503 0.478687
## conference:gender                 1   0.02  0.0158   0.111 0.738859
## student:gender                    1   0.07  0.0741   0.521 0.470905
## format:conference:student         1   0.13  0.1333   0.937 0.333625
## format:conference:gender          1   0.01  0.0076   0.054 0.816947
## format:student:gender             1   0.04  0.0381   0.268 0.605033
## conference:student:gender         1   0.02  0.0205   0.144 0.704415
## format:conference:student:gender  1   0.02  0.0187   0.131 0.717269
## Residuals                       368  52.33  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: 708.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6471 -0.6488  0.0114  0.6185  2.9939
##
## Random effects:
##  Groups     Name        Variance  Std.Dev.  Corr
##  conference (Intercept) 0.000e+00 0.000e+00
##             format1     7.858e-13 8.864e-07   NaN
##             gender1     9.055e-11 9.516e-06   NaN -0.58
##  Residual               1.419e-01 3.767e-01
## Number of obs: 752, groups:  conference, 4
##
## Fixed effects:
##                         Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)             6.092185   0.049535 739.998720 122.987  < 2e-16
## format1                 0.176775   0.051546 739.999773   3.429 0.000638
## student1                0.015552   0.047500 739.999999   0.327 0.743446
## gender1                -0.035621   0.046464 739.987575  -0.767 0.443550
## reviewSingle           -0.017851   0.066458 739.998868  -0.269 0.788313
## time                    0.010594   0.027815 739.997880   0.381 0.703400
## format1:student1        0.053034   0.071131 739.999997   0.746 0.456160
## format1:gender1         0.001137   0.068685 740.000000   0.017 0.986793
## format1:reviewSingle   -0.010324   0.062928 739.999617  -0.164 0.869728
## student1:gender1       -0.041920   0.062783 739.999993  -0.668 0.504542
## student1:reviewSingle  -0.022851   0.061047 739.999999  -0.374 0.708277
## gender1:reviewSingle    0.042268   0.058686 739.980486   0.720 0.471601
##
## (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    -0.133  0.245
## gender1      0.177 -0.228  0.053
## reviewSingl -0.913  0.186  0.049 -0.102
## time        -0.844 -0.032 -0.080 -0.039  0.845
## frmt1:stdn1  0.120 -0.332 -0.483 -0.004 -0.026  0.019
## frmt1:gndr1 -0.155  0.261 -0.017 -0.472  0.066  0.042  0.028
## frmt1:rvwSn  0.196 -0.674 -0.079  0.086 -0.230  0.017  0.019    0.001
## stdnt1:gnd1  0.012  0.016  0.127 -0.223 -0.069 -0.005 -0.079    0.056
## stdnt1:rvwS  0.138 -0.111 -0.633 -0.077 -0.213 -0.011  0.123    0.000
## gndr1:rvwSn -0.079  0.119 -0.059 -0.619  0.146 -0.012 -0.001    0.113
##             frm1:S std1:1 std1:S
## format1
## student1
## gender1
## reviewSingl
## time
## frmt1:stdn1
## frmt1:gndr1
## frmt1:rvwSn
## stdnt1:gnd1  0.007
## stdnt1:rvwS  0.080  0.125
## gndr1:rvwSn -0.107 -0.084  0.000
## 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)
```

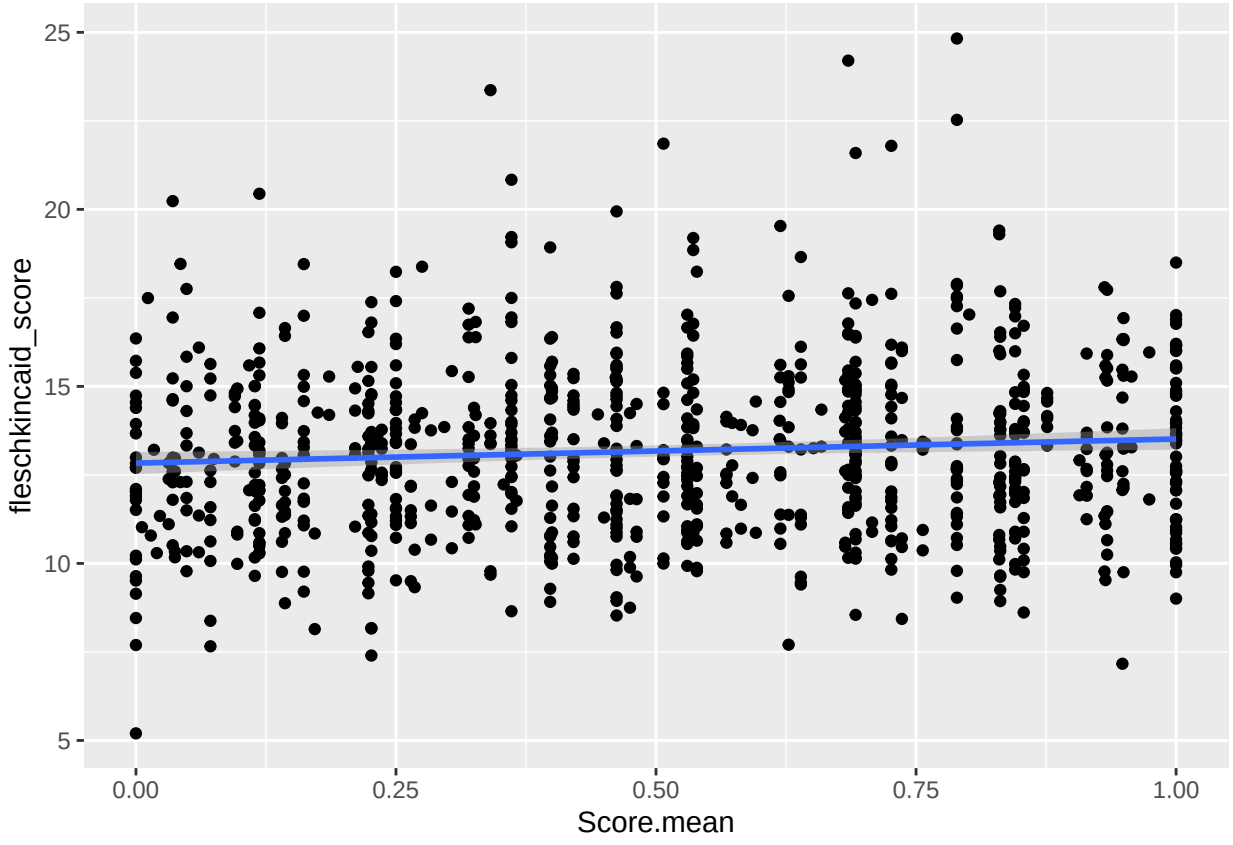
```
##
##  Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$fleschkincaid_score
## t = 2.5333, df = 899, p-value = 0.01147
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01898224 0.14868380
## sample estimates:
```

```
##        cor
## 0.08418961
```

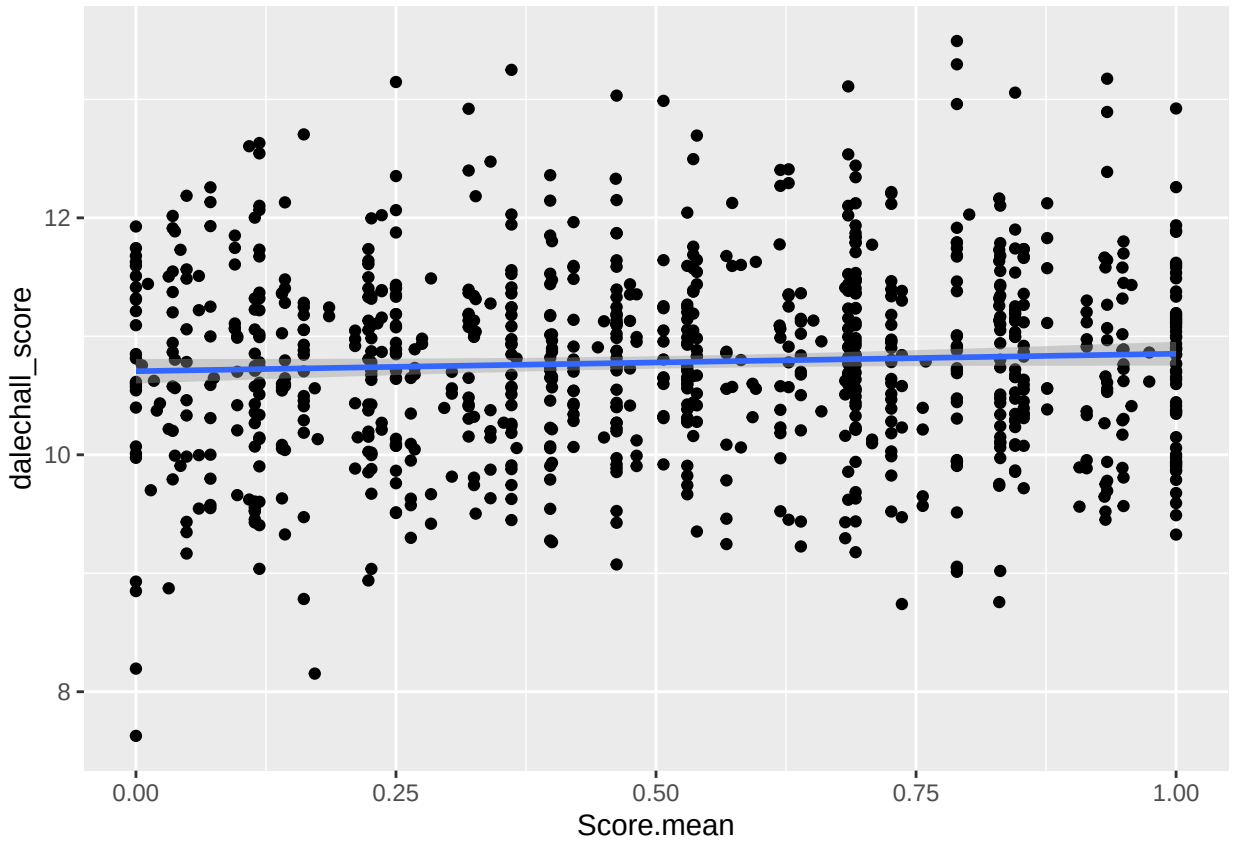
```
cor.test(readScores$Score.mean, readScores$dalechall_score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  readScores$Score.mean and readScores$dalechall_score
## t = 1.6561, df = 899, p-value = 0.09806
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01019975  0.12002814
## sample estimates:
##        cor
## 0.05514873
```

```
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 2123.9 2151.6 -1055.9   2111.9
```

```
## m1  7 2124.5 2156.9 -1055.2   2110.5 1.3612      1     0.2433
## m2  8 2126.5 2163.5 -1055.2   2110.5 0.0128      1     0.9101
```

```
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
##   2126.5   2163.5  -1055.2   2110.5      744
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.91591 -0.89750 -0.01329  0.90172  1.93794
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  conference (Intercept) 0.0000   0.0000
##  Residual               0.9691   0.9844
## Number of obs: 752, groups:  conference, 4
##
## Fixed effects:
##                                    Estimate Std. Error        df t value
## (Intercept)                        -0.03063    0.04543 752.00000  -0.674
## format1                             0.24444    0.08280 752.00000   2.952
## student1                           -0.01572    0.07812 752.00000  -0.201
## gender1                             0.12435    0.07538 752.00000   1.650
## fleschkincaid_score_scaled          0.04664    0.04139 752.00000   1.127
## gender1:fleschkincaid_score_scaled  0.00930    0.08233 752.00000   0.113
##                                   Pr(>|t|)
## (Intercept)                        0.50037
## format1                            0.00325 **
## student1                           0.84056
## gender1                            0.09945 .
## fleschkincaid_score_scaled         0.26018
## gender1:fleschkincaid_score_scaled 0.91009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) formt1 stdnt1 gendr1 flsc__
## format1     -0.469
## student1    -0.359  0.088
## gender1      0.272 -0.115  0.024
## flschkncd__  0.026 -0.133  0.011 -0.019
## gndr1:fls__ -0.077  0.062  0.064 -0.041  0.368
## 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) +
## m1:     dalechall_score_scaled
## m2: Score.mean.norm ~ format + student + gender + (1 | conference) +
## m2:     dalechall_score_scaled + gender:dalechall_score_scaled
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0  6 2123.9 2151.6 -1055.9   2111.9
## m1  7 2125.8 2158.2 -1055.9   2111.8 0.0126      1     0.9105
## m2  8 2127.8 2164.8 -1055.9   2111.8 0.0015      1     0.9695
```

```
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
##   2127.8   2164.8  -1055.9   2111.8      744
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.88762 -0.90250 -0.01729  0.90949  1.89758
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  conference (Intercept) 0.0000   0.0000
##  Residual               0.9709   0.9853
## Number of obs: 752, groups:  conference, 4
##
## Fixed effects:
##                                Estimate Std. Error         df t value
## (Intercept)                   -0.032914   0.045398 752.000000  -0.725
## format1                        0.257878   0.083635 752.000000   3.083
## student1                      -0.015425   0.078086 752.000000  -0.198
## gender1                        0.125463   0.075510 752.000000   1.662
```

```
## dalechall_score_scaled           0.004044   0.038333 752.000000   0.105
## gender1:dalechall_score_scaled  -0.002873   0.075164 752.000000  -0.038
##                                Pr(>|t|)
## (Intercept)                     0.46868
## format1                         0.00212 **
## student1                        0.84345
## gender1                         0.09702 .
## dalechall_score_scaled          0.91601
## gender1:dalechall_score_scaled  0.96952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) formt1 stdnt1 gendr1 dlch__
## format1     -0.467
## student1    -0.358  0.089
## gender1      0.273 -0.120  0.024
## dlchll_scr_  0.061 -0.205 -0.026  0.015
## gndr1:dlc__ -0.040  0.055  0.025 -0.052  0.147
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

No interactions.