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)
library(stargazer)
```

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

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

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

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

Look at the distribution of submissions:

```
table(allData$FirstAuthorGender,allData$conference)
```

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

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

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

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

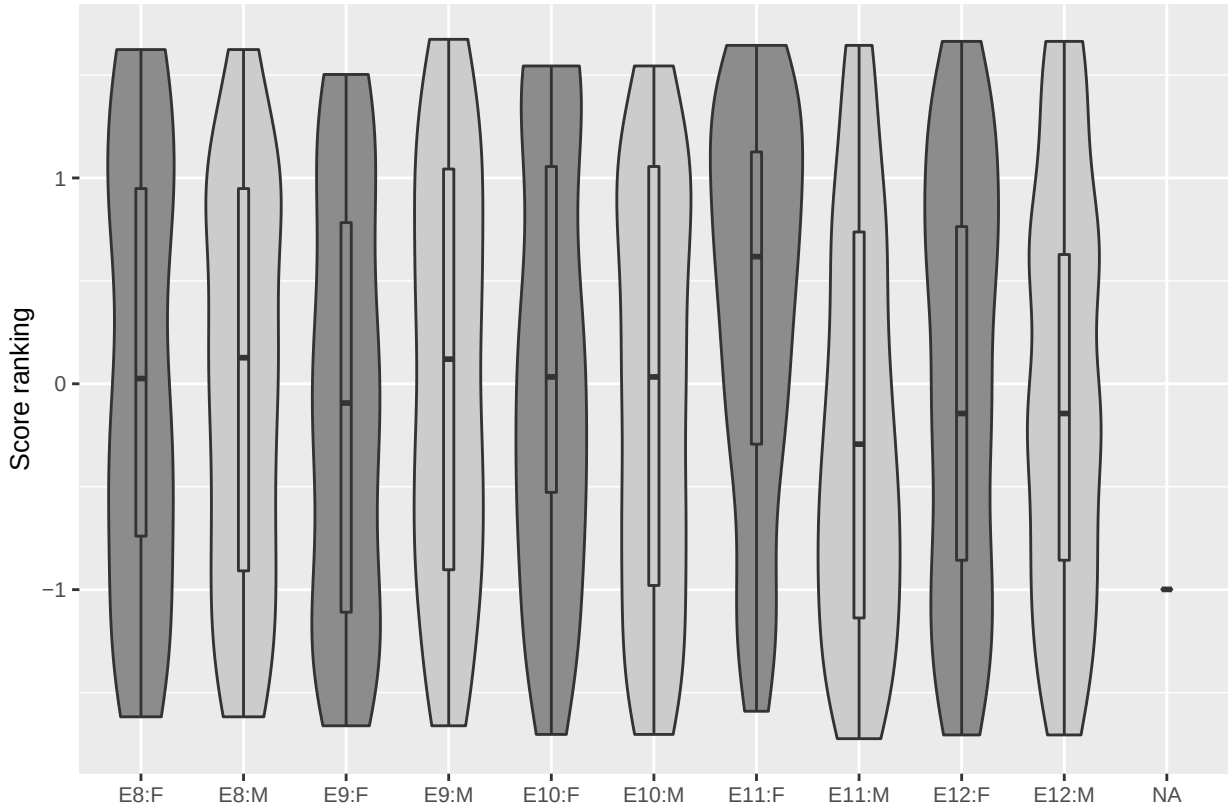
```
## , ,  = Non-Student
##
##
##     E8 E9 E10 E11 E12
##   F  0 34  55  41  54
##   M  0 85  94  77  93
##
## , ,  = Student
##
##
##     E8 E9 E10 E11 E12
##   F  0 18  12  35  30
```

```
##   M  0 45  30  42  29
```

Plots

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

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

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

p2
```

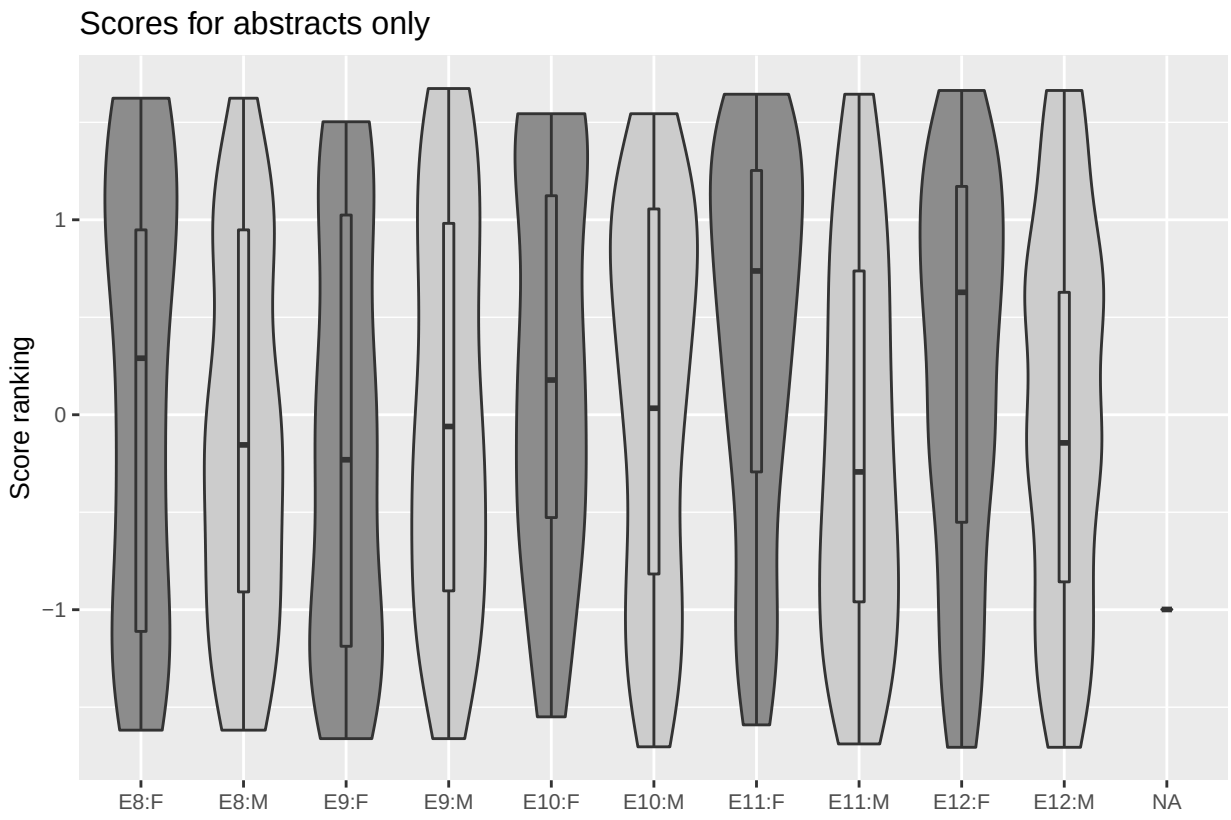



```
pdf("../results/Results_Gender_3conf.pdf", width = 12, height= 6)
p2
dev.off()
```

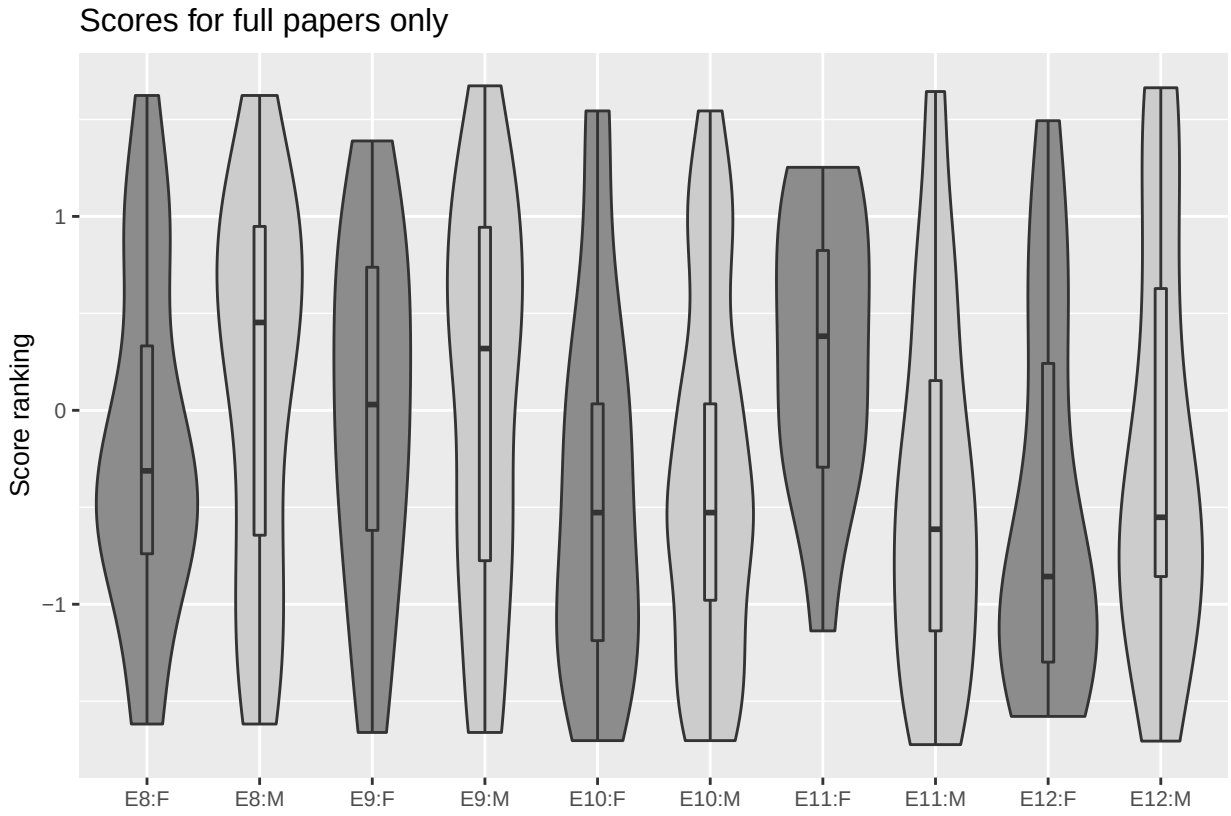
```
## pdf
##   2
```

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

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

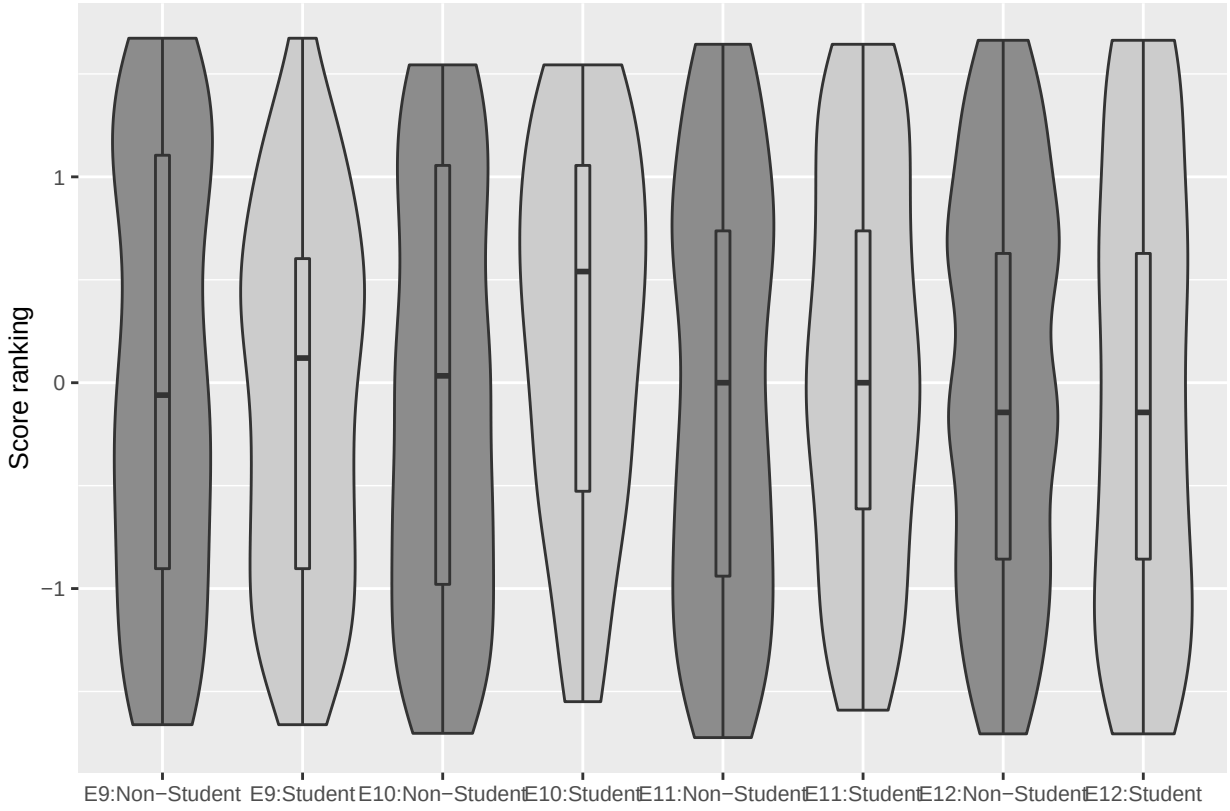



```
p2Paper <- ggplot(allData[allData$format=="Paper",],
             aes((conference):(FirstAuthorGender), Score.mean,
                 fill=FirstAuthorGender))

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



Rank by student status in each conference.

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

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

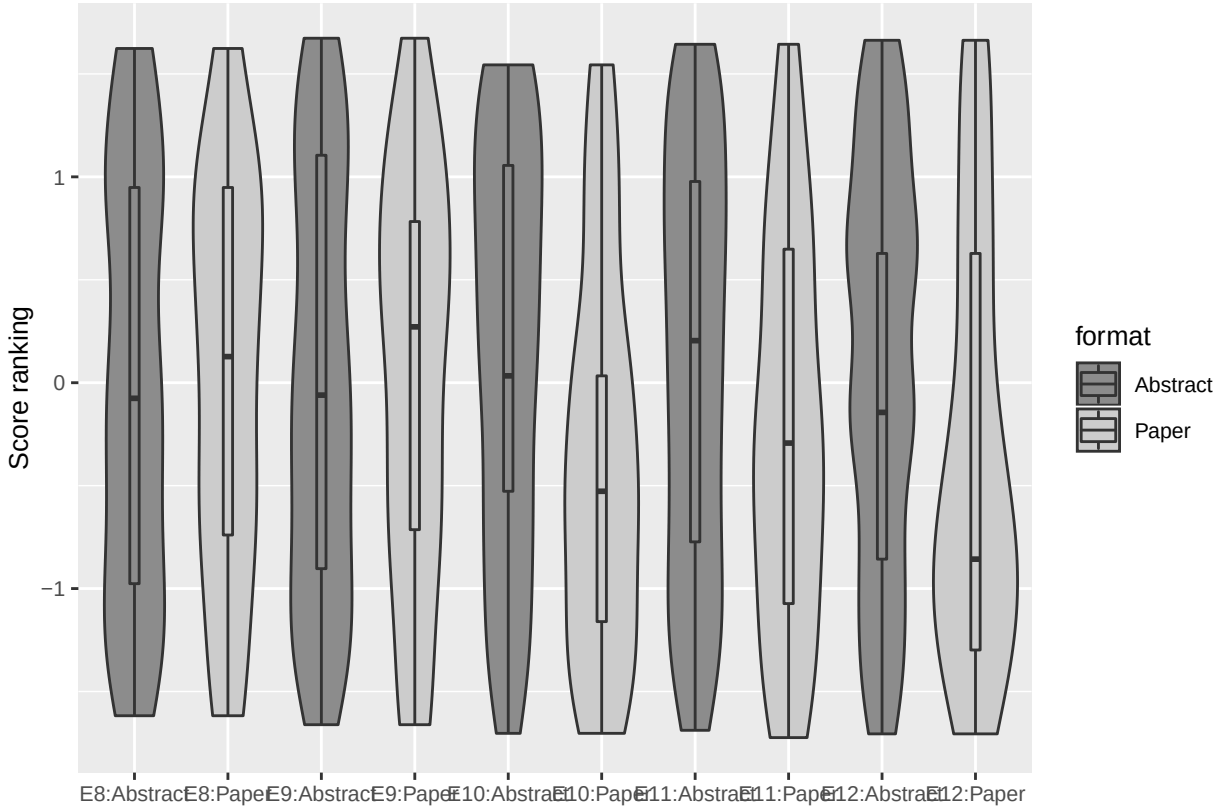



```
pdf("../results/Results_Student_3conf.pdf", width = 12, height= 6)
p
dev.off()
```

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

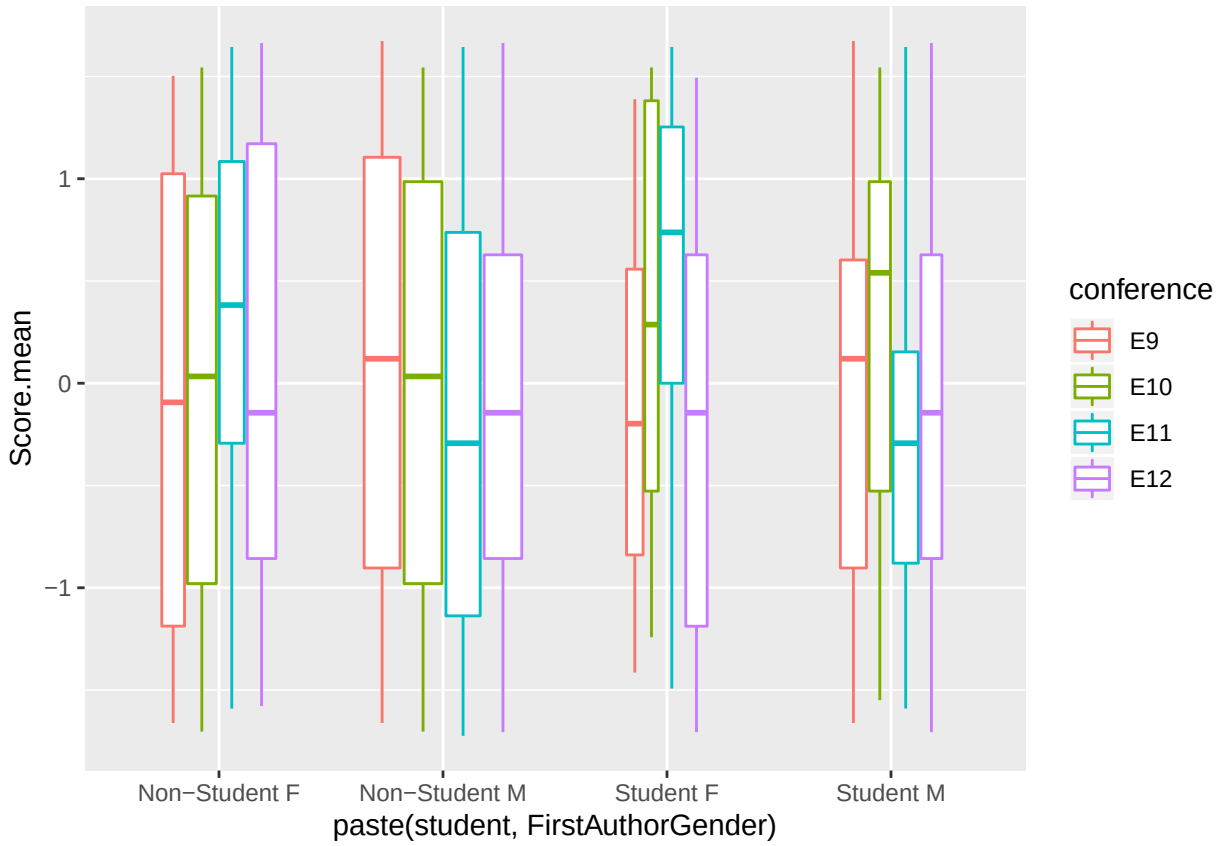
Format:

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

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

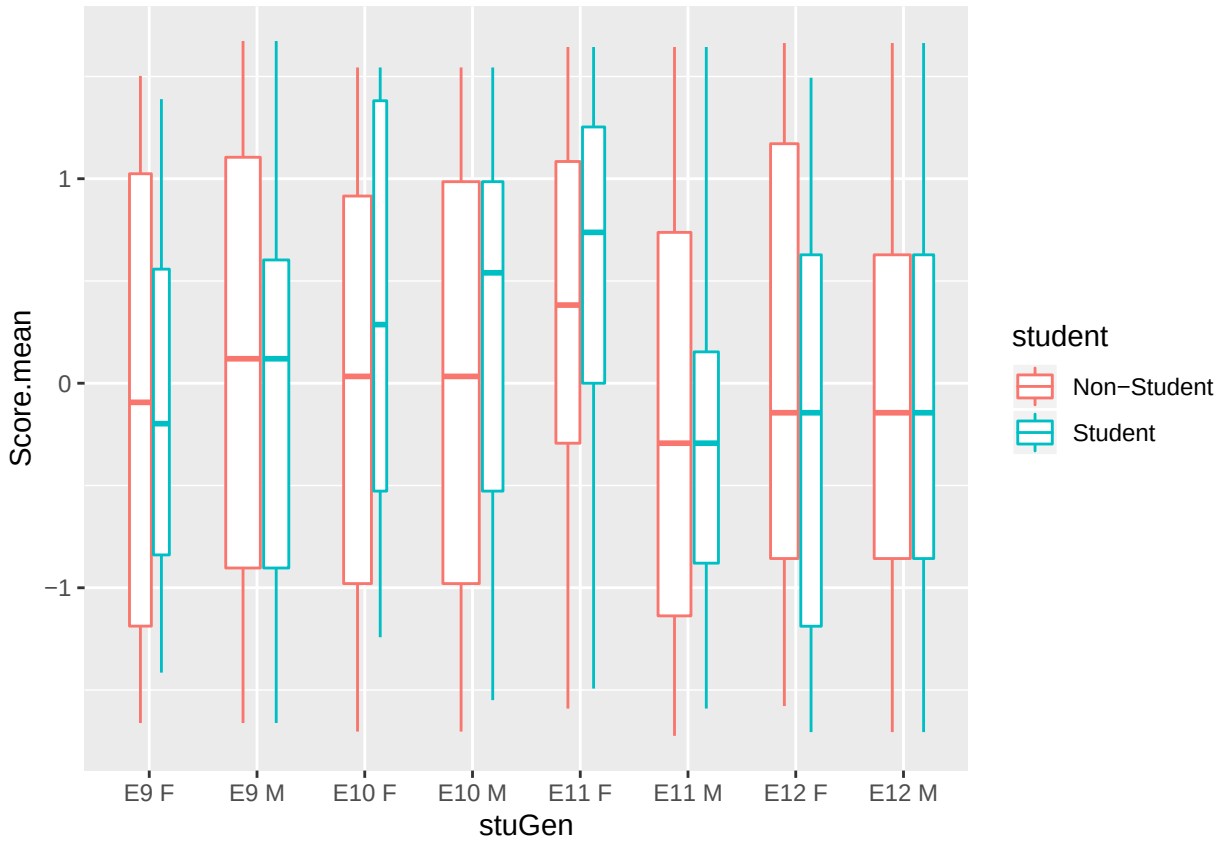


Combined student and gender:

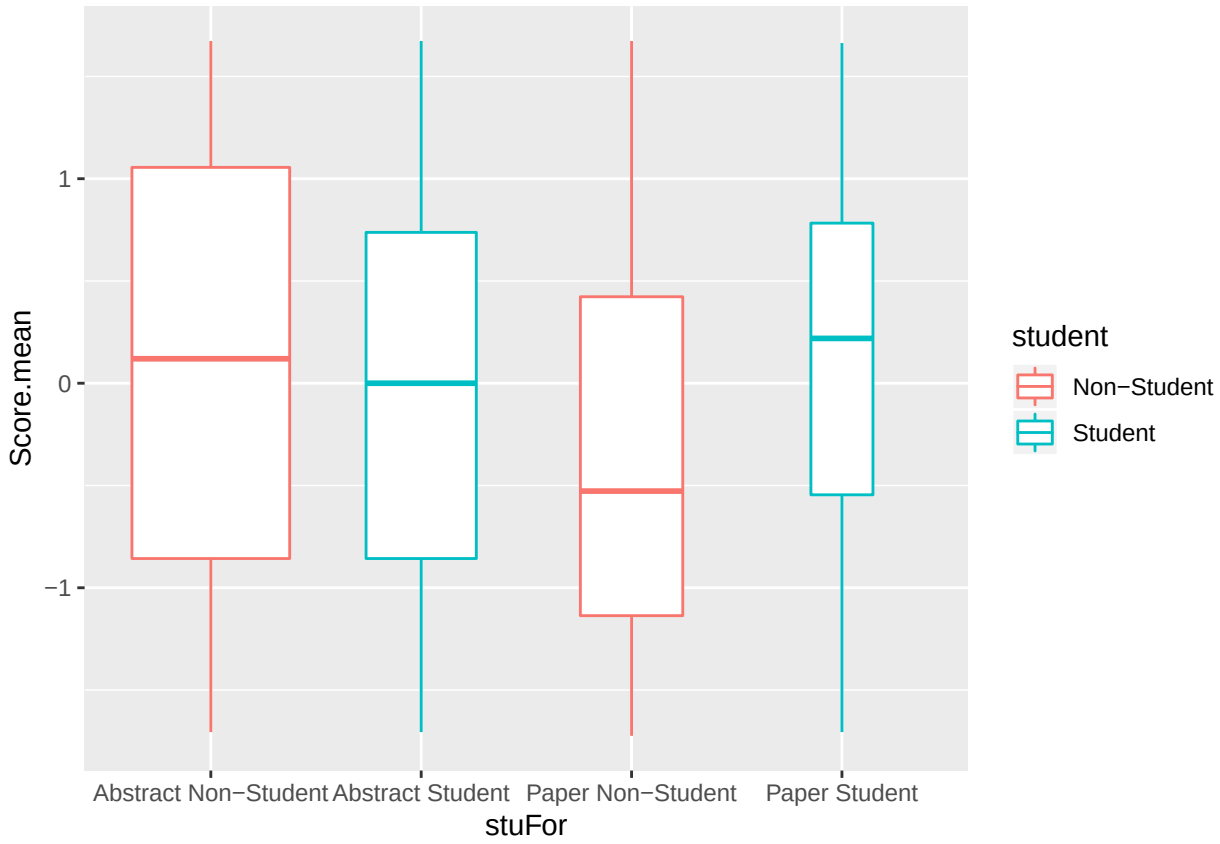
```
ggplot(allData[allData$conference!="E8",],
       aes(y=Score.mean,x=paste(student,FirstAuthorGender),colour=conference))+ geom_boxplot(varwidth =
```




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

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

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




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

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



Summary statistics

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

cbind(t1,t2,t3)
```

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

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

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

```
# Statistics for female authors:
stargazer(allData[allData$FirstAuthorGender=="F",],type = 'text')
```

```
##
```

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

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

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

Review ranks by gender and student status

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

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

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

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

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

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

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

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

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

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

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

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

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

Mixed effects model

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

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

```
summary(m0)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula:
## Score.mean ~ 1 + (FirstAuthorGender * review * student * format) +
##     (0 + FirstAuthorGender + student + format | conference)
##    Data: allData[allData$conference != "E8", ]
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+07))
##
## REML criterion at convergence: 2175.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0469 -0.8318 -0.0649  0.8731  2.1003
##
## Random effects:
##  Groups     Name                Variance Std.Dev. Corr
##  conference FirstAuthorGenderF 0.049878 0.22333
##             FirstAuthorGenderM 0.002765 0.05258  -0.97
##             student1           0.045642 0.21364  -0.87  0.73
##             format1            0.023844 0.15441   0.37 -0.14 -0.77
##  Residual                      0.950489 0.97493
## Number of obs: 774, groups:  conference, 4
##
## Fixed effects:
##                                          Estimate Std. Error
## (Intercept)                             -0.005526   0.063844
## FirstAuthorGender1                       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)                                    2.851857  -0.087   0.9367
## FirstAuthorGender1                             2.601224   0.882   0.4519
## review1                                        2.851857  -0.738   0.5163
## student1                                       2.859042  -1.428   0.2528
## format1                                        3.396149   1.269   0.2845
## FirstAuthorGender1:review1                     2.601224   0.771   0.5046
## FirstAuthorGender1:student1                  674.800702  -1.101   0.2714
## review1:student1                               2.859042   0.762   0.5040
## FirstAuthorGender1:format1                   749.272361   0.467   0.6405
## review1:format1                                3.396149   1.178   0.3148
## student1:format1                             615.328521   3.278   0.0011 **
## FirstAuthorGender1:review1:student1          674.800701   0.186   0.8526
## FirstAuthorGender1:review1:format1           749.272361   0.474   0.6357
## FirstAuthorGender1:student1:format1          719.938801   0.663   0.5077
## review1:student1:format1                     615.328521  -1.434   0.1521
## FirstAuthorGender1:review1:student1:format1 719.938801   0.200   0.8414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

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

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

```
m0 <- lmer(
      Score.mean ~
        1 + (FirstAuthorGender*review*format) +
        (0+FirstAuthorGender+format|conference),
      allData,
  control=lmerControl(optimizer="bobyqa",optCtrl = list(maxfun=10000000)),
  REML = T
)
```

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

```
summary(m0)
```

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

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

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

Permutation test

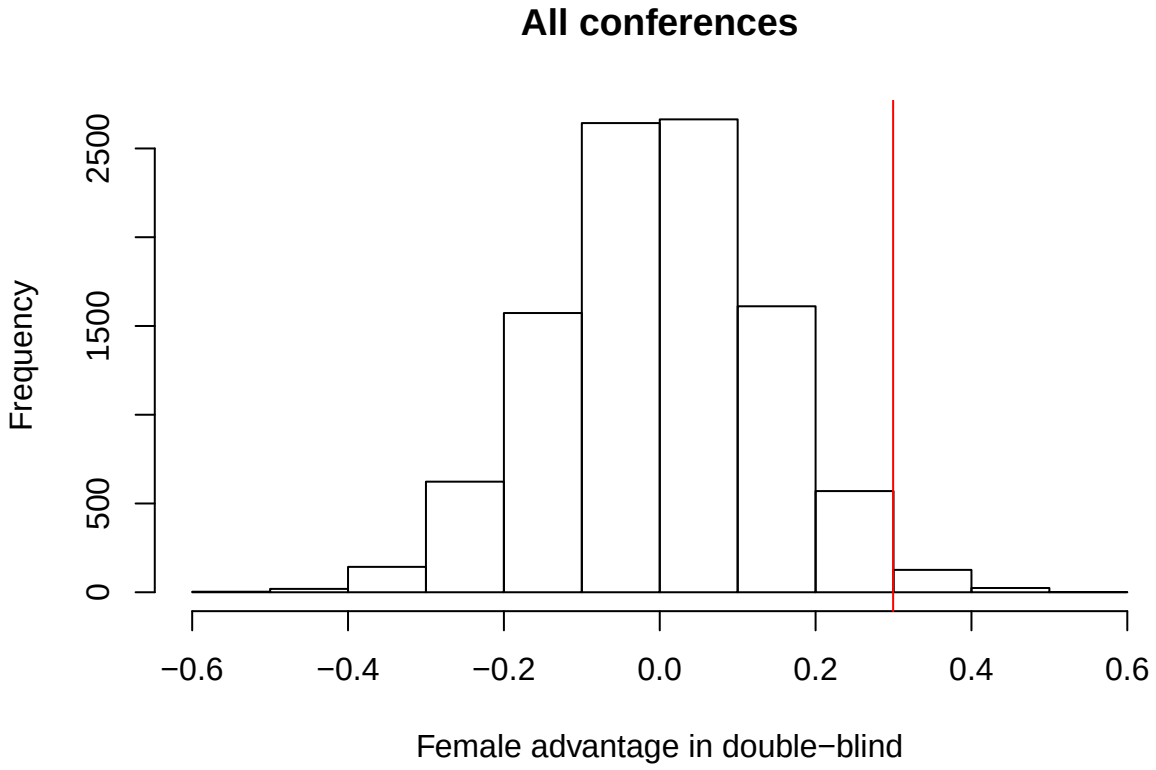
The distributions of score means are not very normal within conferences. We run a permutation test to address this. We calculate the average difference between single blind and double blind scores for males (dM) and for females (dF). Then we calculate dF - dM. A value > 0 means females scores increase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

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

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

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

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

Permutation test for all data:

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

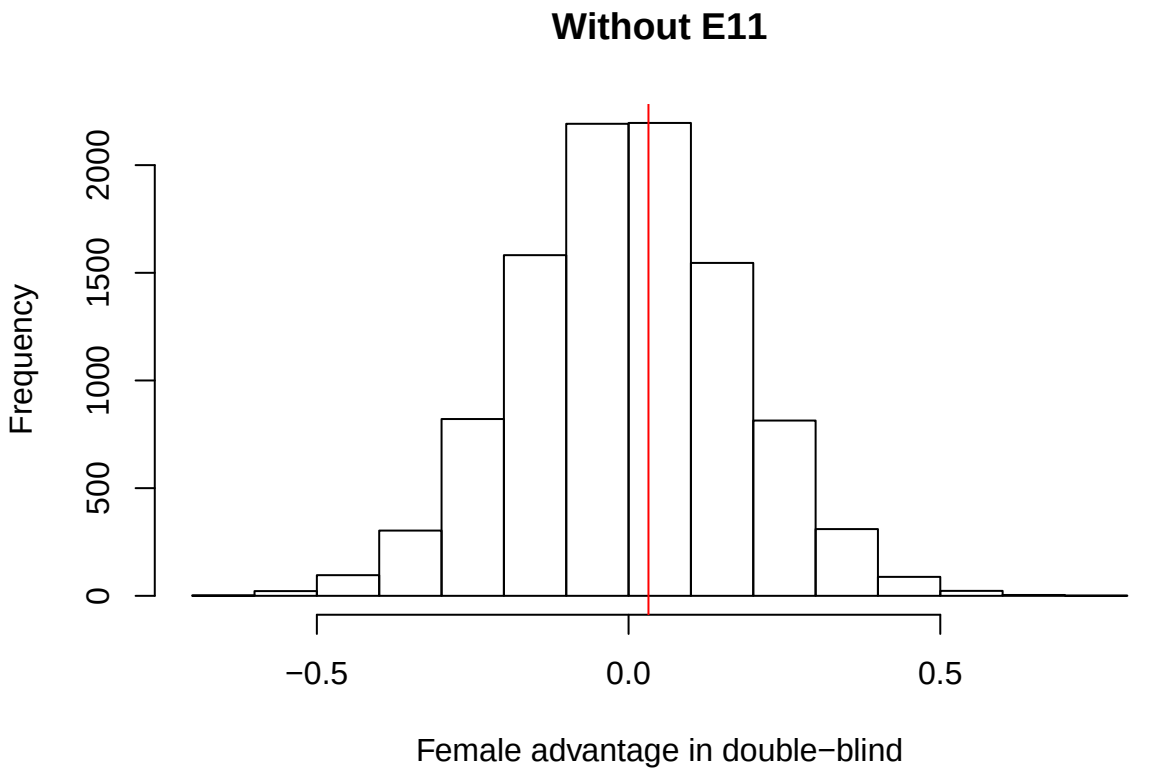
```
## [1] "p= 0.0152 , z= 2.15468786786684"
```



Permutation test without E11 data:

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

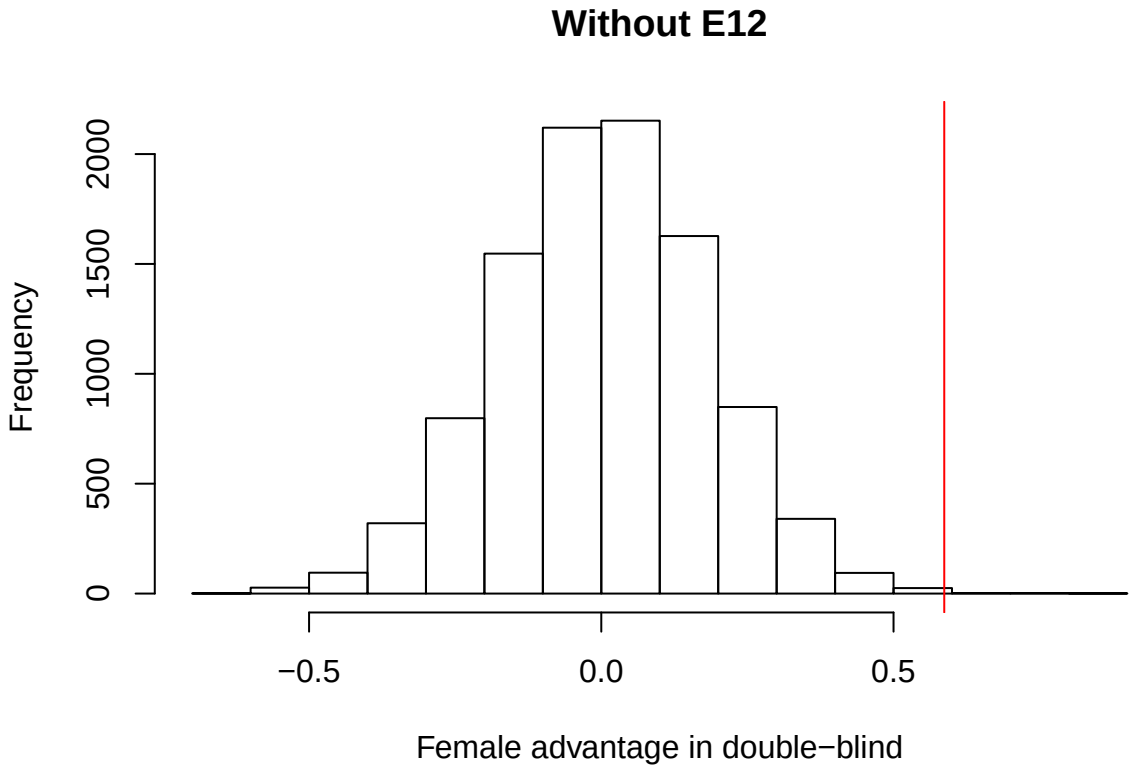
```
## [1] "p= 0.4209 , z= 0.185895800347046"
```



Permutation test without E12 data:

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

```
## [1] "p= 7e-04 , z= 3.28946384764505"
```

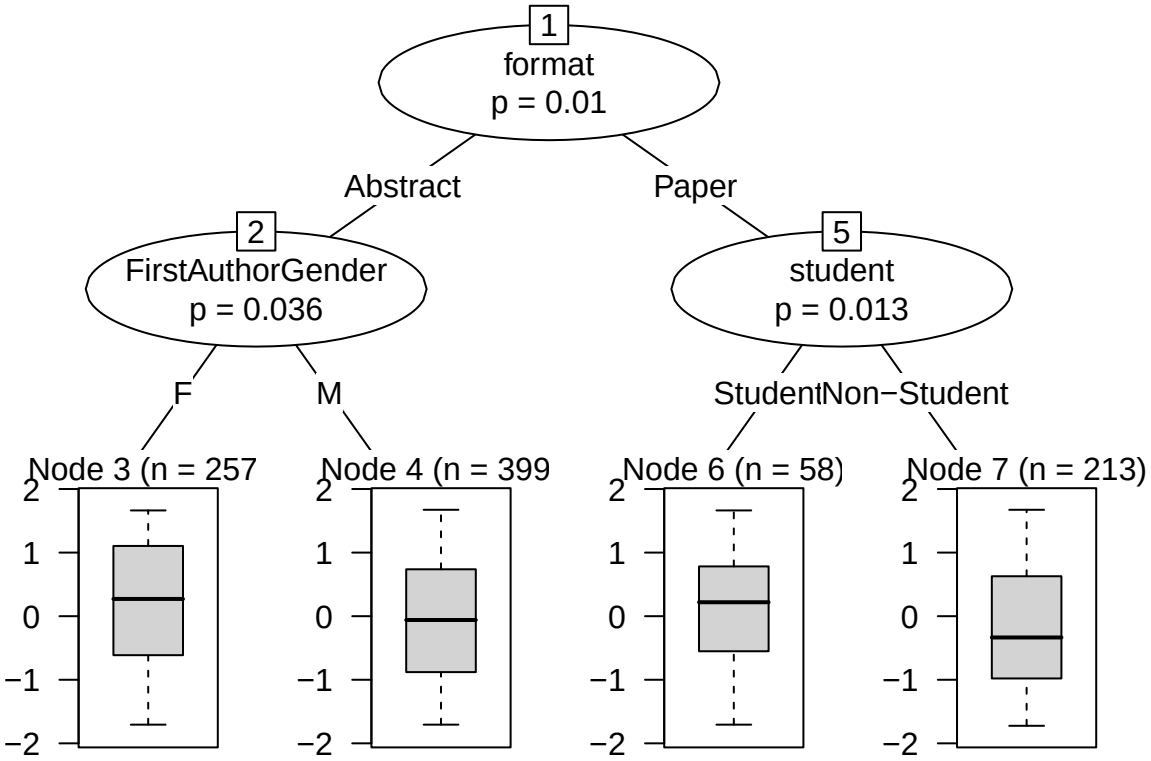


The results are in line with the test above. Across the whole data, females are given higher scores in double-blind, but this is driven by E11 alone.

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).

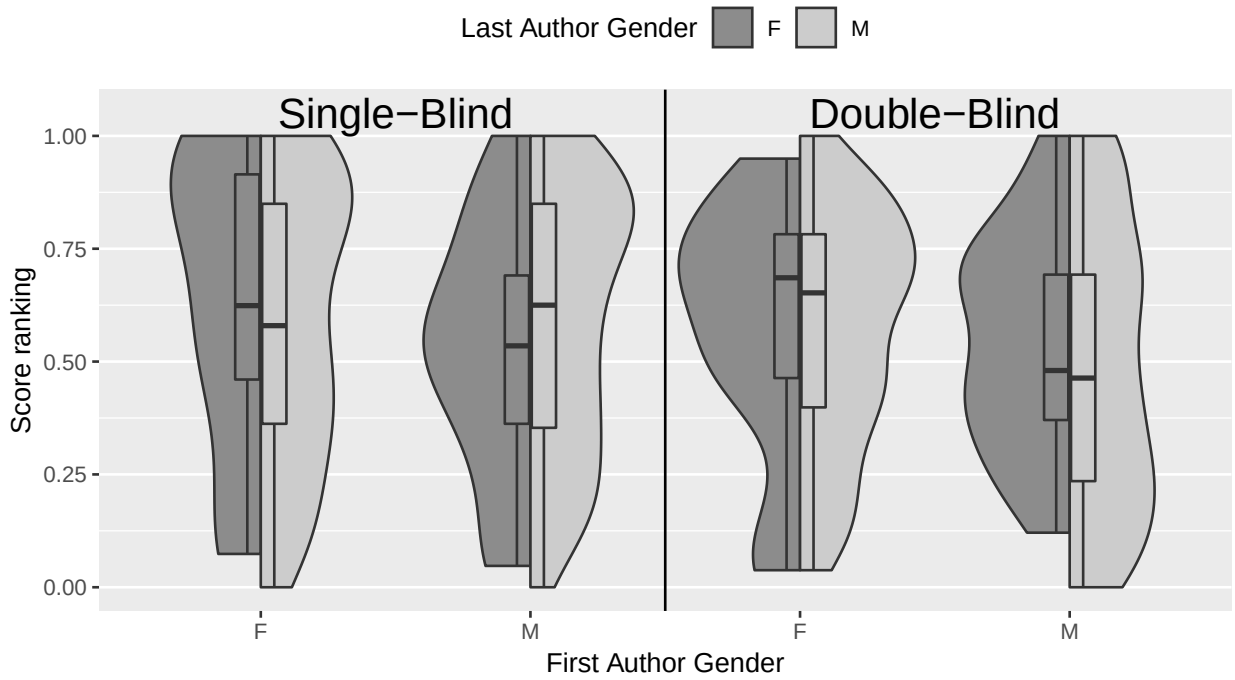


Figure 1: Distributions of review scores by first and last author gender.

Influence of last author

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

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

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

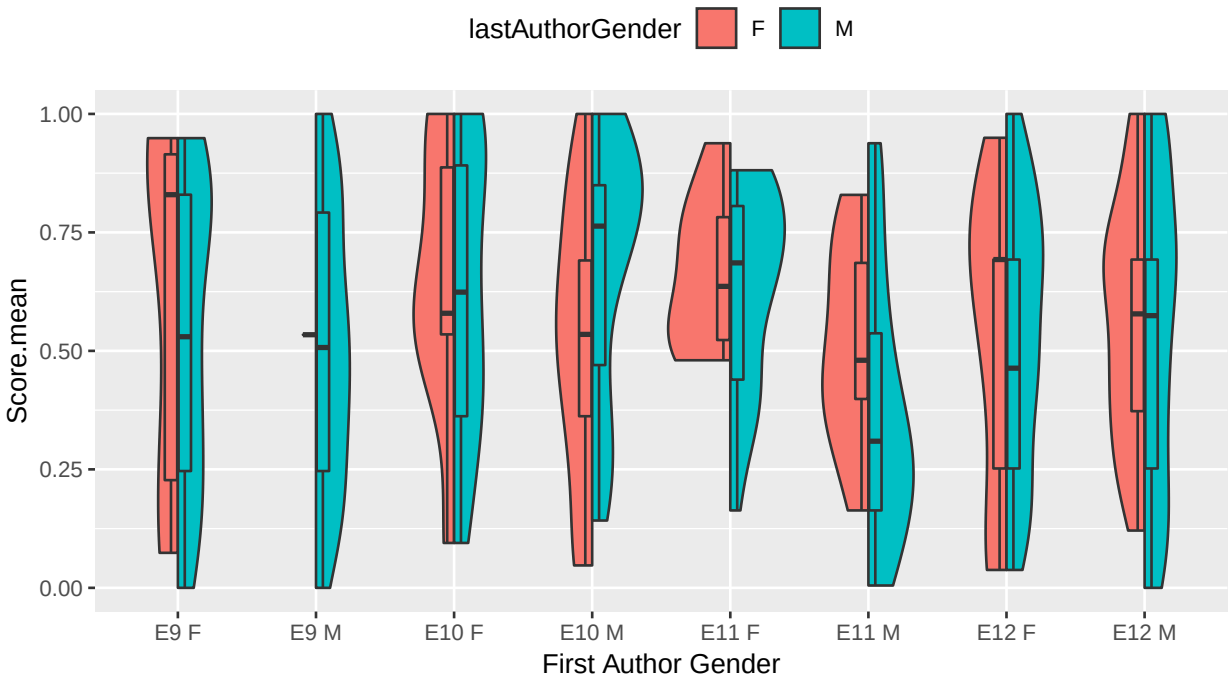


Figure 2: Distributions of review scores by first and last author gender.

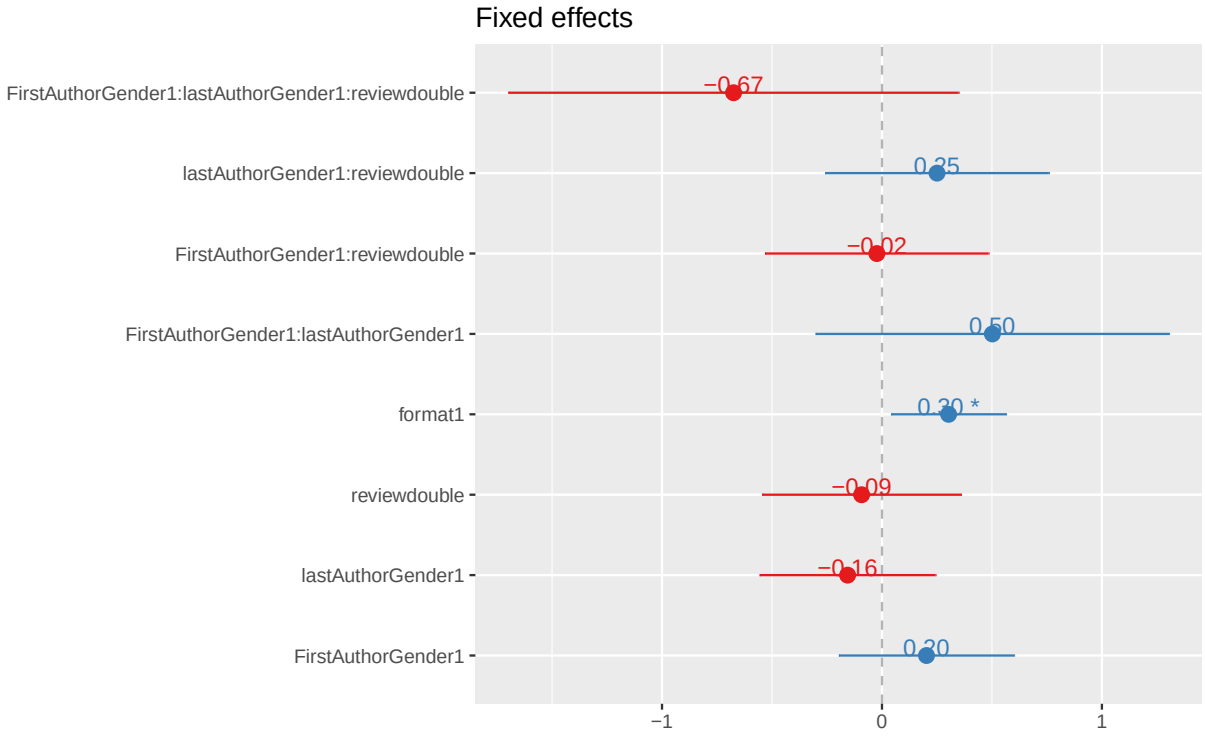


Figure 3: Coefficients and confidence intervals for effects predicting review ranks.