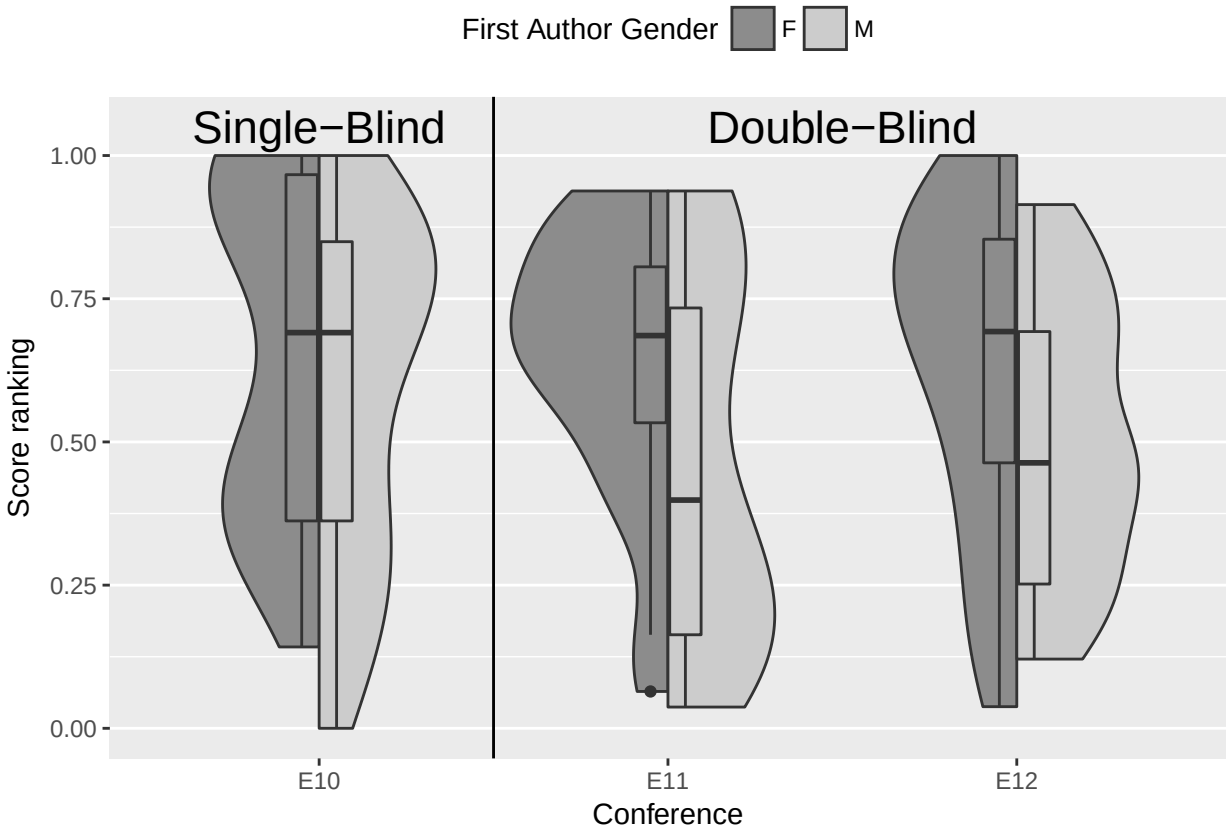
Matched sample E10, E11 and E12

Load libraries and extra plotting functions.

```
library(ggplot2)
library(lme4)
source("SplitViolinPlot.R")
library(tidyr)
library(lavaan)
library(semPlot)
library (scales)
```

```
allC = read.csv("../data/MatchedAuthors_E10_E11_E12.csv")
allC.long = allC %>% gather(conference,Score.mean,2:4)
allC.long$format = allC.long$E10.format
allC.long[allC.long$conference=="E11",]$format =
  allC.long[allC.long$conference=="E11",]$E11.format
allC.long[allC.long$conference=="E12",]$format =
  allC.long[allC.long$conference=="E11",]$E12.format
```

```
gx = ggplot(allC.long,
       aes(conference, Score.mean, fill=gender)) +
  annotate("text", x = c(1,2.5), y = c(1.05,1.05),
           label=c("Single-Blind", "Double-Blind"), size=6) +
  geom_split_violin() +
  geom_vline(xintercept=1.5) +
  scale_y_continuous(name="Score ranking", breaks = c(0,0.25,0.5,0.75,1))+
  scale_x_discrete(name="Conference")+
  scale_fill_grey(start = 0.55, end=0.8,name="First Author Gender") +
  geom_boxplot(width=0.2, show.legend = F) +
  theme(legend.position = "top",
        panel.grid.major.x = element_blank())
gx
```




```
pdf("../results/MatchedSamples.pdf",
    height=5,width=6)
gx
dev.off()
```

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

Fit a mixed effects model with random intercepts for author. The key question is whether there is an interaction between gender and review Type.

```
contrasts(allC.long$gender) <- contr.sum(2)/2
contrasts(allC.long$format) <- contr.sum(2)/2
allC.long$reviewType =
  as.factor(c("Single","Double")[
    1+(allC.long$conference %in% c("E11","E12"))])
contrasts(allC.long$reviewType) <- contr.sum(2)/2

m0 = lmer(scale(Score.mean)~ 1 +
       format +
       (1|authorCode),
        data = allC.long)
# Check if review type needs a random slope
mRevRan = update(m0,~.+(0+reviewType|authorCode))
anova(m0,mRevRan)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: allC.long
## Models:
## m0: scale(Score.mean) ~ 1 + format + (1 | authorCode)
## mRevRan: scale(Score.mean) ~ format + (1 | authorCode) + (0 + reviewType |
## mRevRan:     authorCode)
##         Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0       4 409.60 421.64 -200.8   401.60
## mRevRan  7 415.41 436.48 -200.7   401.41 0.1892      3     0.9793
```

```
#No

# Add variales
mGen = update(m0,~.+gender)
anova(m0,mGen)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: allC.long
## Models:
## m0: scale(Score.mean) ~ 1 + format + (1 | authorCode)
## mGen: scale(Score.mean) ~ format + (1 | authorCode) + gender
##      Df   AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0    4 409.6 421.64 -200.8    401.6
## mGen  5 407.2 422.26 -198.6    397.2 4.3912      1    0.03613 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mRev = update(mGen,~.+reviewType)
anova(mGen,mRev)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: allC.long
## Models:
## mGen: scale(Score.mean) ~ format + (1 | authorCode) + gender
## mRev: scale(Score.mean) ~ format + (1 | authorCode) + gender + reviewType
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mGen  5 407.20 422.26 -198.60   397.20
```

```
## mRev  6 405.82 423.88 -196.91   393.82 3.3866      1    0.06573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mGenxRev = update(mRev,~.+gender:reviewType)
anova(mRev,mGenxRev)
```

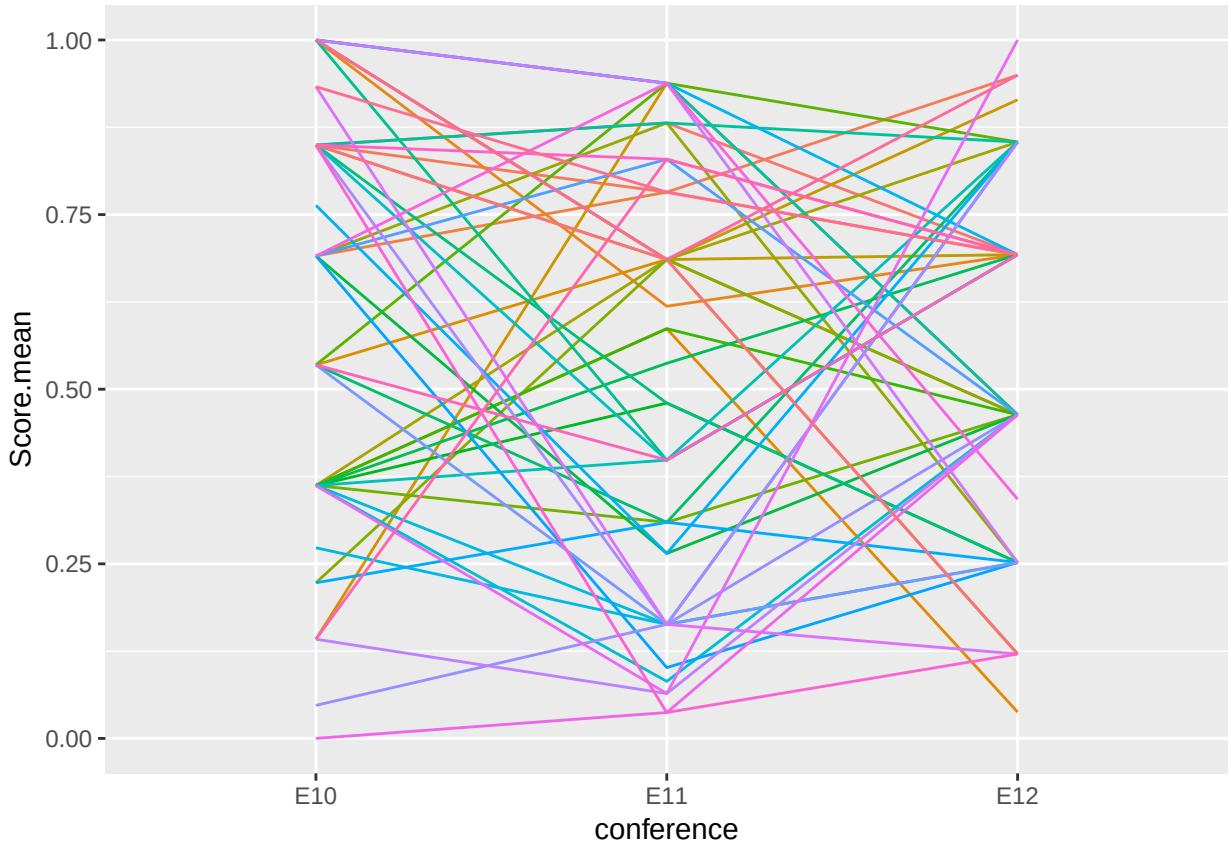
```
## refitting model(s) with ML (instead of REML)
```

```
## Data: allC.long
## Models:
## mRev: scale(Score.mean) ~ format + (1 | authorCode) + gender + reviewType
## mGenxRev: scale(Score.mean) ~ format + (1 | authorCode) + gender + reviewType +
## mGenxRev:     gender:reviewType
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mRev      6 405.82 423.88 -196.91   393.82
## mGenxRev  7 405.70 426.78 -195.85   391.70 2.1146      1     0.1459
```

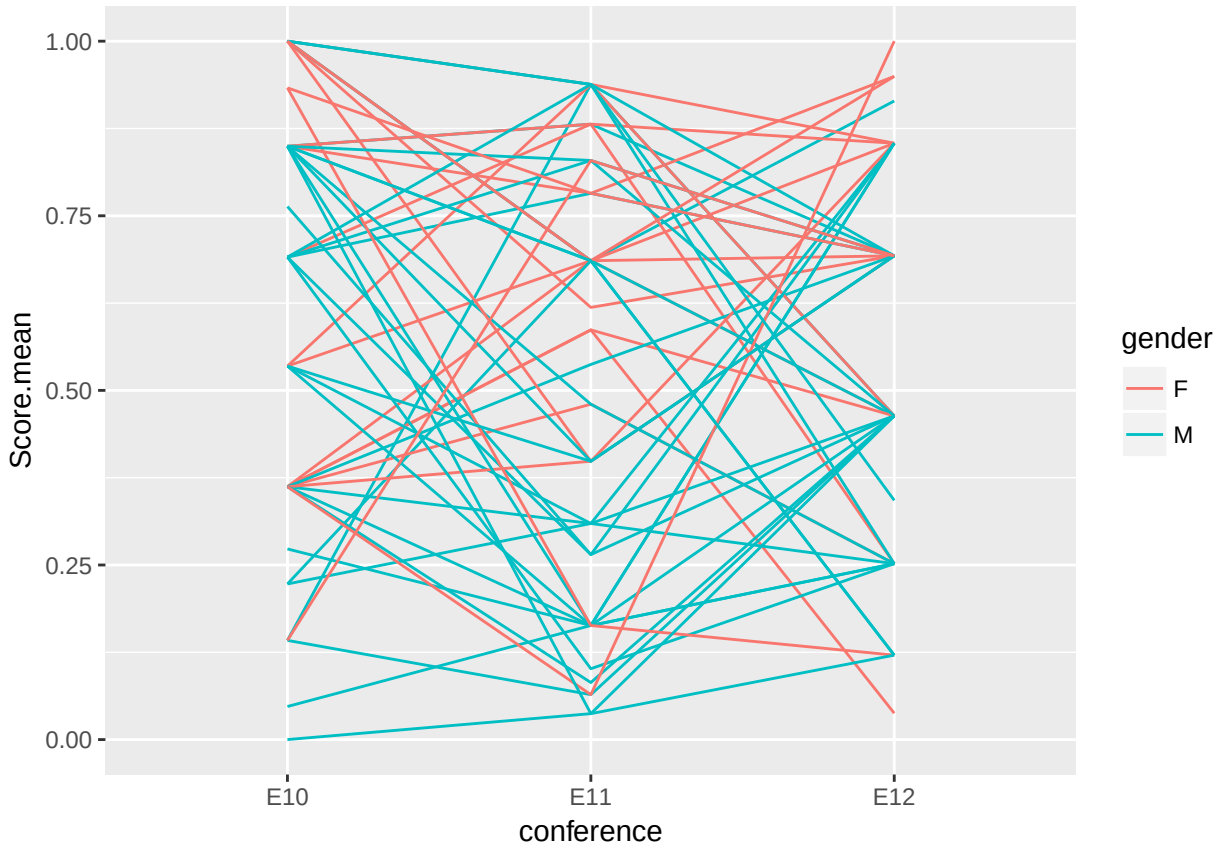
Significant effect of gender (in this sample, female authors tend to receive better scores overall than males). As in the full sample, abstracts receive higher scores than papers. There is no significant interaction between gender and review type.

Plot individual data:

```
ggplot(allC.long,
       aes(y=Score.mean,
           x=conference,
           group=authorCode,
           colour=authorCode)) +
  geom_line() + theme(legend.position = 'none')
```

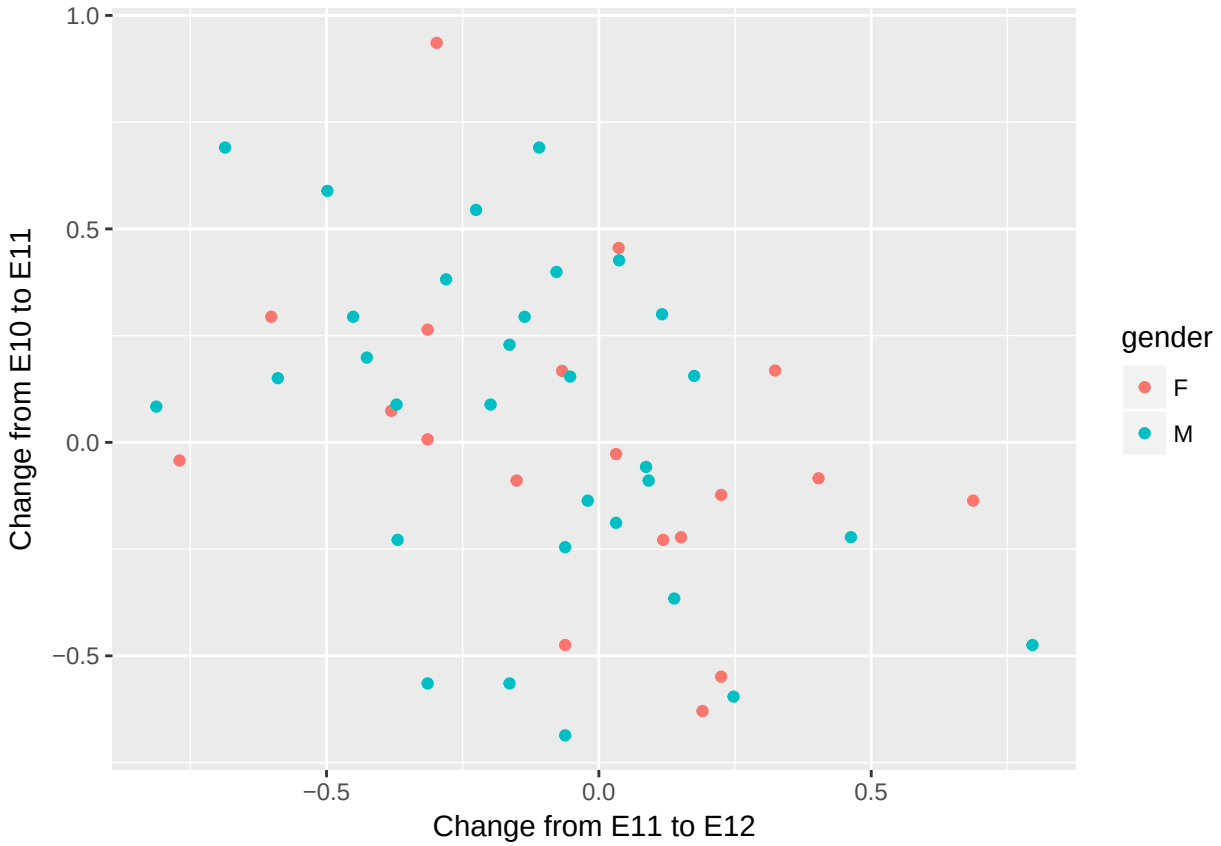



```
ggplot(allC.long,
       aes(y=Score.mean,
           x=conference,
           group=authorCode,
           colour=gender)) +
  geom_line()
```



Plot improvement from E10 to E11 against improvement from E11 to E12.

```
ggplot(allC, aes(x=diff.E10.to.E11,y=diff.E11.to.E12,colour=gender)) + geom_point() +
  xlab("Change from E11 to E12") +
  ylab("Change from E10 to E11")
```



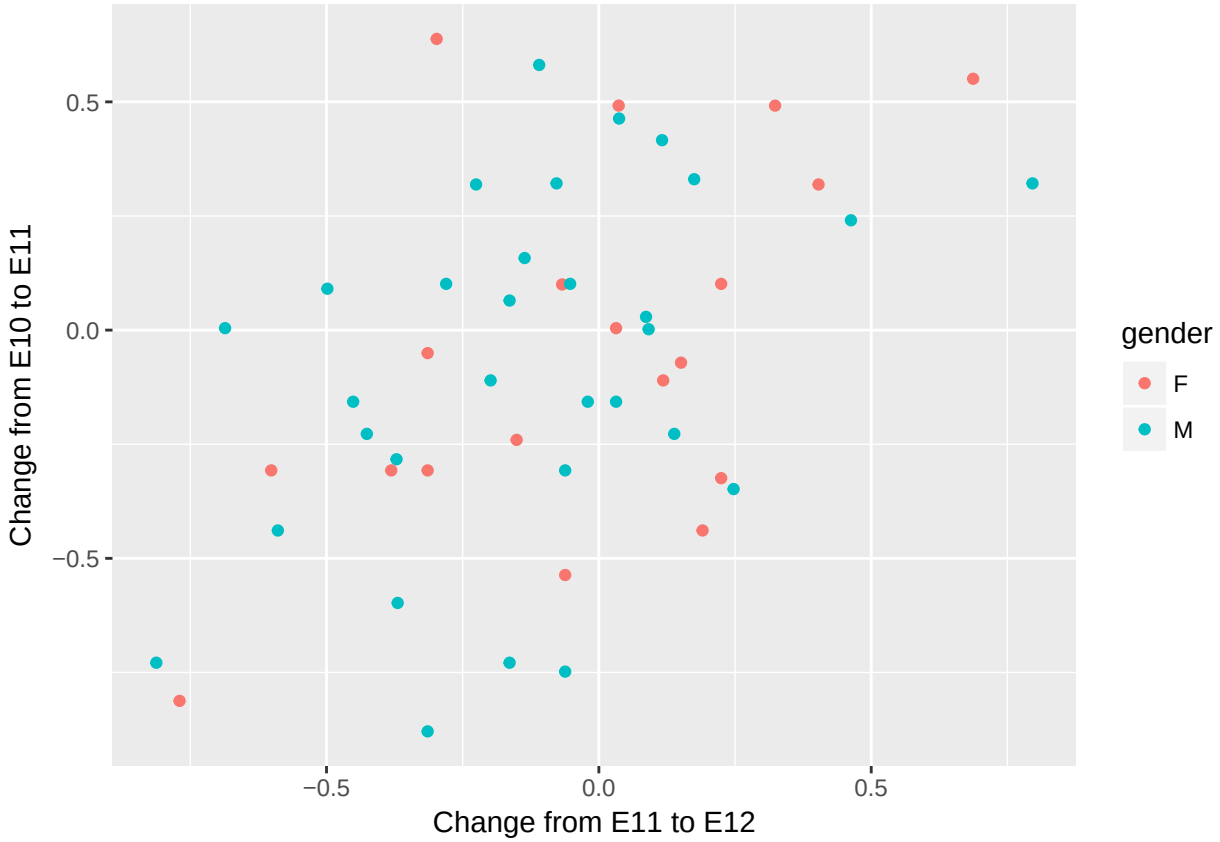
This seems to suggest that researchers who improved from E10 to E11 tended to do worse going from E11 to E12. There seems to be no effect of gender on this relationship:

```
contrasts(allC$gender) <- contr.sum(2)/2
summary(lm(diff.E11.to.E12~diff.E10.to.E11*gender,
           data=allC))
```

```
##
## Call:
## lm(formula = diff.E11.to.E12 ~ diff.E10.to.E11 * gender, data = allC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70367 -0.15658  0.00494  0.22562  0.84414
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.022350   0.052881  -0.423  0.67453
## diff.E10.to.E11          -0.447880   0.151422  -2.958  0.00488 **
## gender1                  -0.003911   0.105762  -0.037  0.97066
## diff.E10.to.E11:gender1   0.117924   0.302843   0.389  0.69879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3523 on 46 degrees of freedom
## Multiple R-squared:  0.1765, Adjusted R-squared:  0.1228
## F-statistic: 3.286 on 3 and 46 DF,  p-value: 0.02893
```

There's also a positive relationship between improvement from E10 to E11 and the difference between E12 and E10. i.e. if you improved from E10 to E11, then you will have improved from E10 to E12.

```
ggplot(allC, aes(x=diff.E10.to.E11,y=E12-E10,colour=gender)) + geom_point() +
  xlab("Change from E11 to E12") +
  ylab("Change from E10 to E11")
```



Only 10% of researchers improve year on year:

```
table((allC$E10 < allC$E11) & (allC$E11 < allC$E12))
```

```
##
## FALSE  TRUE
##    45     5
```

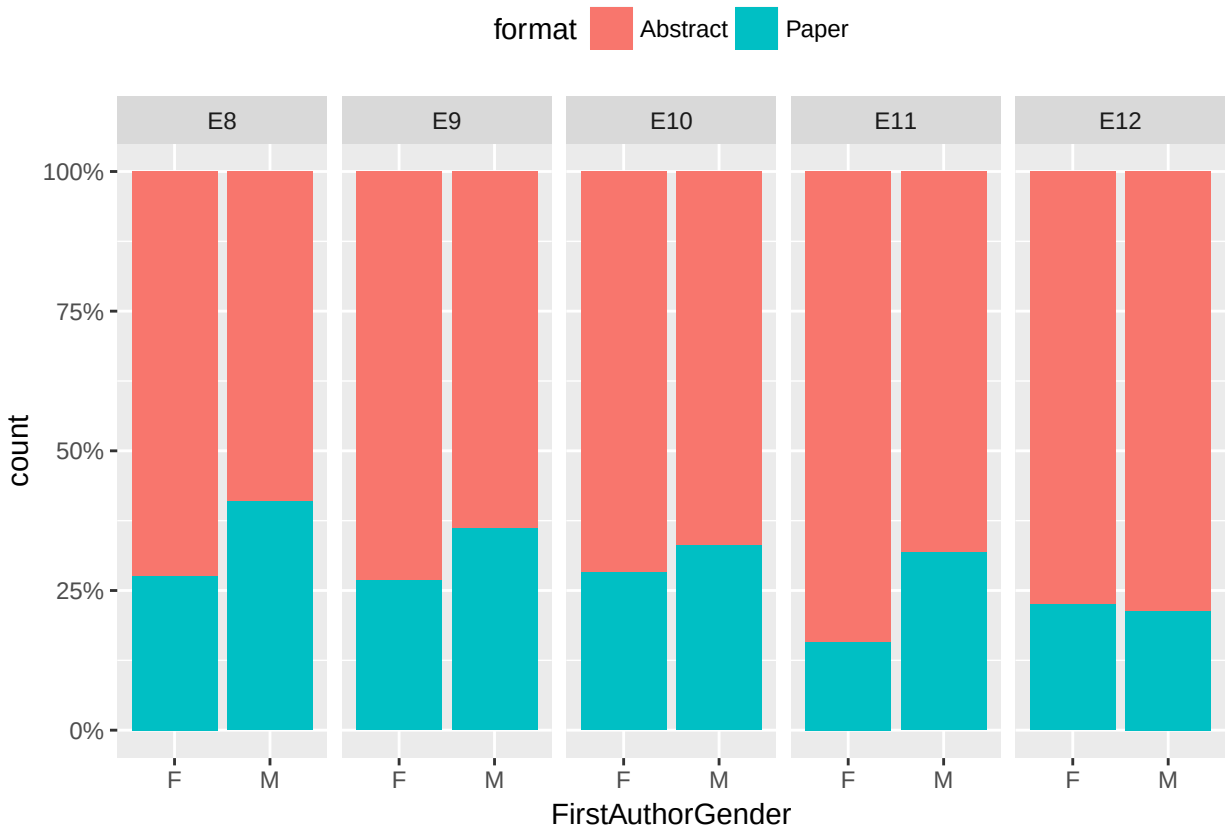
Causal approach

In the analysis below, we use a structural equation model based on the hypothesised causal structure of the data.

Part of the reason for doing this is that submission format varies with gender. Because abstracts are more likely to be given higher scores than papers in general, this might be confounding the effect of gender. Below is a graph created from the whole data (not just the matched samples):

```
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)

ggplot(allData, aes(FirstAuthorGender)) +
    geom_bar(aes(fill=format), position = "fill") +
  scale_y_continuous(labels = percent_format())+
  facet_grid(~conference) +
  theme(legend.position = "top")
```



From the raw distribution, we see that male authors are more likely to submit a paper than female authors (though this trend varies by conference).

In the structural equation model, we assume that:

- The author's score for one conference has an impact on the score in the following year.
- The author's choice of submission format in a given year influences the score (independently for that year).
- The author's gender influences which format they submit.
- The author's gender affects the score in E10 (single blind review)
- The author's gender does not affect the score in E11 nor E12 (double blind reivew), but we estimate a correlation.

```
allC$gender.F = allC$gender=="F"
allC$E10.abstract = allC$E10.format=="Abstract"
allC$E11.abstract = allC$E11.format=="Abstract"
allC$E12.abstract = allC$E12.format=="Abstract"
model = "
E12 ~ E11
E11 ~ E10
E10 ~ NA*E10.abstract + gender.F
E11 ~ E11.abstract
E11 ~~ gender.F
E12 ~ E12.abstract
E12 ~~ gender.F
E10.abstract ~ gender.F
E11.abstract ~ gender.F
E12.abstract ~ gender.F
"
fit <- sem(model, data=allC)
summary(fit, standardized=TRUE)
```

```
## lavaan (0.5-23.1097) converged normally after  42 iterations
##
##   Number of observations                            50
##
##   Estimator                                         ML
##   Minimum Function Test Statistic               25.456
##   Degrees of freedom                                10
##   P-value (Chi-square)                           0.005
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Standard Errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   E12 ~
##     E11               0.012    0.118    0.102    0.919    0.012    0.013
##   E11 ~
##     E10               0.243    0.127    1.917    0.055    0.243    0.247
##   E10 ~
##     E10.abstract      0.280    0.088    3.167    0.002    0.280    0.430
##     gender.F         -0.007    0.083   -0.085    0.932   -0.007   -0.012
##   E11 ~
##     E11.abstract      0.117    0.080    1.467    0.142    0.117    0.189
##   E12 ~
```

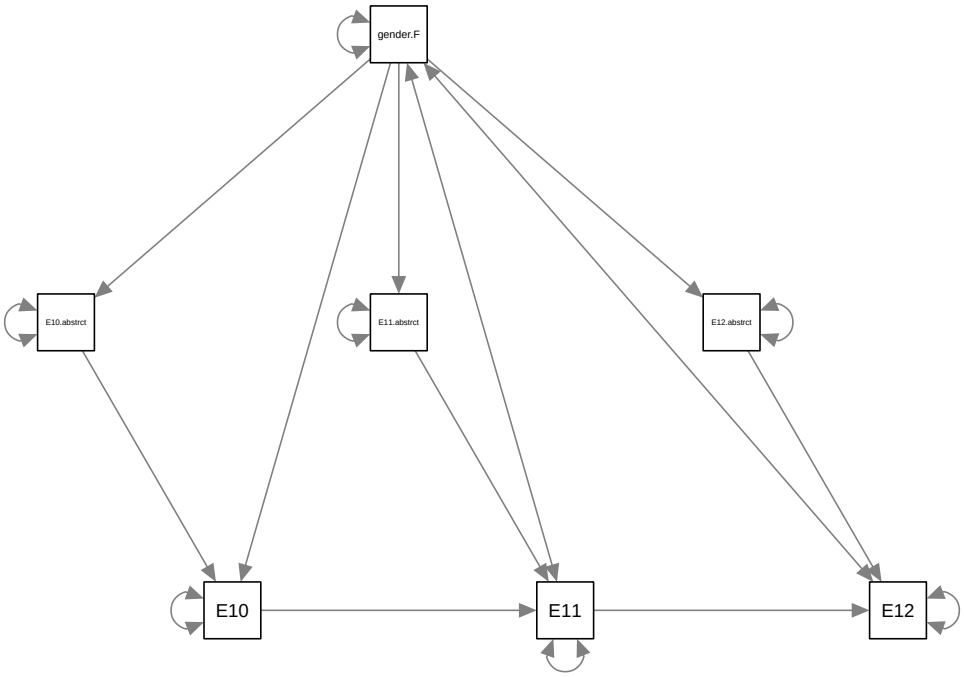
```
##     E12.abstract      0.264    0.080    3.296    0.001    0.264    0.413
##   E10.abstract ~
##     gender.F          0.314    0.126    2.496    0.013    0.314    0.333
##   E11.abstract ~
##     gender.F          0.124    0.137    0.906    0.365    0.124    0.127
##   E12.abstract ~
##     gender.F         -0.070    0.120   -0.579    0.563   -0.070   -0.082
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  .E11 ~~
##     gender.F          0.034    0.020    1.716    0.086    0.034    0.253
##  .E12 ~~
##     gender.F          0.032    0.017    1.919    0.055    0.032    0.274
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .E12               0.060    0.012    4.997    0.000    0.060    0.846
##    .E11               0.075    0.015    4.990    0.000    0.075    0.875
##    .E10               0.073    0.015    5.000    0.000    0.073    0.818
##    .E10.abstract      0.187    0.037    5.000    0.000    0.187    0.889
##    .E11.abstract      0.221    0.044    5.000    0.000    0.221    0.984
##    .E12.abstract      0.170    0.034    5.000    0.000    0.170    0.993
##     gender.F          0.236    0.047    5.002    0.000    0.236    1.000
```

Plots:

```
layout = matrix(c(
  1,6,  1,4,  1,2,  2,1,  2,3,  2,5,  3,3),
  ncol=2,byrow = T)
layout = layout[,2:1]
```

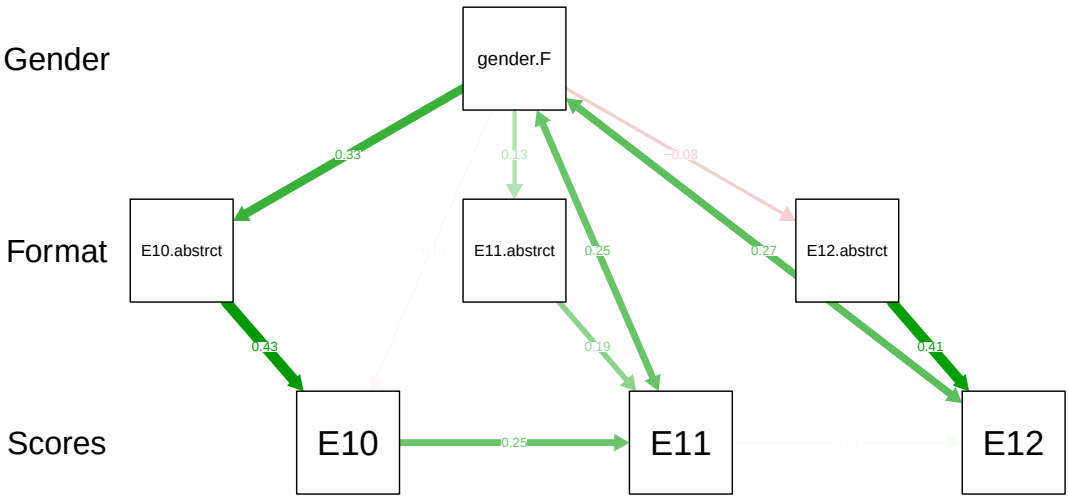
Model:

```
semPaths(fit,nCharNodes = 11,layout=layout)
```



Model with standard estimates:

```
semPaths(fit, 'std',layout=layout,residuals=F,intercepts=F,exoVar=F,
         exoCov=T,nCharNodes=11,sizeMan=10,shapeMan="rectangle")
text(-1.3,-1,"Scores")
text(-1.3,0,"Format")
text(-1.3,1,"Gender")
```



The direct effect of gender on score for E10 is not different from zero (the distributions for male and female authors is similar for E10). The relationship between gender and score for E11 and E12 is positive (female authored papers receive higher scores than male authored papers), but only marginally significant.

The effect of gender on format is strong for E10 (female authors prefer to submit abstracts), but weaker for E11 and E12. The effect of format on score is significant for E10 and E12 (abstracts score higher than papers), but weaker not for E11.

Improvement from E10 to E11 is marginally significant, but there is no improvement from E11 to E12. The correlations between reported above may be due to colliders in the causal graph.