Li_esli_analysis

Anna Artemova

28 January 2020

Data preprocessing

```
# find and list all the files starting with "L".
# Maybe add full regualar expr for the files

filenames = list.files(pattern="^[L]")

myfiles = lapply(filenames, read_csv)

# add a variable for filename
myfiles <- Map(cbind, myfiles, from = filenames)

for (i in 1:length(myfiles)) {
  myfiles[[i]]$word_count <- as.numeric(myfiles[[i]]$word_count)
}

library(stringr)

# an empty object for the loop
data <- list()

# first, create the list variable in each list,
# then take all list to merge them in one

# subject_nr! in 27 need to change to 27
# subject_nr! in 38 need to change to 38

for (i in 1:length(filenames)) {
  myfiles[[i]]$subj_id <- str_sub(filenames[i], 1, 5)
  myfiles[[i]]$subject_nr <- str_sub(filenames[i], 2, 3)

  data <- bind_rows(data, myfiles[i])
}

# create a dataset
data <- data %>%
  select(subj_id,
         subject_nr,
         avg_rt,
         coding,
         condition,
         conditon,
         count_exp_sequence,
         main_asp,
         main_num,
         main_tense,
         num,
         response,
         response_time,
         response_time_next,
         sentence_rus,
         sub_asp,
         sub_num,
         sub_tense,
         sub_wo,
         subject_parity,
         time_exp_sequence,
         time_next,
         time_pause,
         time_response_1,
         word_count,
         from)

# there was a typo in the randomization table, that's why we have condition and
# conditon as variables in the dataset

data %>%
  mutate(condition = ifelse(is.na(conditon) == T, condition, conditon)) %>%
  select(-conditon) %>%
  mutate(word_count = ifelse(num == 42, 6, word_count)) -> data

# data is clean
```

```
# read file with demografic data

demogr <- read.csv('demographics.csv',sep = ';' ) %>%
  mutate(subj_id = as.factor(subj_id))

working_data <- data %>% mutate(subj_id = as.factor(subj_id)) %>%
  full_join(demogr, by = 'subj_id') %>% filter(exclude == "no", english == 1) %>%
  mutate(heritageness = droplevels(heritageness))%>%
  mutate(
    subj_id = as.factor(subj_id),
    subject_nr = as.factor(subject_nr),
    coding = as.factor(coding),
    condition = as.factor(condition),
    main_asp = as.factor(main_asp),
    main_num = as.factor(main_num),
    main_tense = as.factor(main_tense),
    response = as.character(response),
    sub_asp = as.factor(sub_asp),
    sub_num = as.factor(sub_num),
    sub_tense = as.factor(sub_tense),
    sub_wo = as.factor(sub_wo),
    from = as.factor(from)
  ) %>%
  filter(coding != c("10distr", "2li", "2distr", #the training trials
                     "12esli", "12li"))

# descriptive stats
```

```
demogr %>%
  filter(exclude == "no", english == 1) %>%
  mutate(female = case_when(gender == "male" ~ 0,
                            gender == "female" ~ 1))%>%
  group_by(heritageness, age_group) %>%
  summarise(count_n = n(),
            mean_age=mean(age, na.rm = T),
            max_age=max(age, na.rm = T),
            min_age=min(age, na.rm = T),
            sd_age = sd(age, na.rm = T),
            female = sum(female, na.rm = T),
            mean_edu=mean(edu_years, na.rm = T),
            english = sum(english, na.rm = T),
            french = sum(french, na.rm = T),
            ukranian = sum(ukranian, na.rm = T),
            chinese = sum(chinese, na.rm = T),
            italian = sum(italian, na.rm = T),
            japanese = sum(japanese, na.rm = T),
            spanish = sum(spanish, na.rm = T),
            hebrew = sum(hebrew, na.rm = T),
            german = sum(german, na.rm = T),
            georgian = sum(georgian, na.rm = T)) -> demogr_summary

demogr_summary %>% select(heritageness, age_group, count_n, mean_age, min_age, max_age, female, mean_edu
```

```
## # A tibble: 4 x 8
## # Groups:   heritageness [2]
##   heritageness age_group count_n mean_age min_age max_age female mean_edu
##   <fct>        <fct>       <int>    <dbl>   <int>   <int>  <dbl>    <dbl>
## 1 no           old             5     50.8      42      74      3     16.3
## 2 no           young          15     22.3      18      31     12     15.6
## 3 yes          old            12     50.2      42      69      5     18.1
## 4 yes          young           5     24.4      20      29      1     13.9
```

```
demogr_summary %>% select(heritageness, age_group, count_n, english, french, ukranian, chinese, italian
```

```
## # A tibble: 4 x 13
## # Groups:   heritageness [2]
##   heritageness age_group count_n english french ukranian chinese italian
##   <fct>        <fct>       <int>   <int>  <int>    <int>   <int>   <int>
## 1 no           old             5       5      1        0       0       1
## 2 no           young          15      15      2        0       1       0
## 3 yes          old            12      12      4        2       1       0
## 4 yes          young           5       5      3        0       0       0
## # ... with 5 more variables: japanese <int>, spanish <int>, hebrew <int>,
## #   german <int>, georgian <int>
```

```
demogr %>%
  filter(exclude == "no") %>%
  filter(heritageness == "yes") %>%
  select(heritageness, age_group, Immigr_data_the_USA) %>%
  mutate(Immigr_data_the_USA = as.numeric(as.character(Immigr_data_the_USA)),
         years_inUSA = 2019 - Immigr_data_the_USA) %>%
  group_by(heritageness, age_group) %>%
  summarise(count_n = n(),
            mean_y_InUSA = mean(years_inUSA),
            min_y_InUSA = min(years_inUSA),
            max_y_InUSA = max(years_inUSA))
```

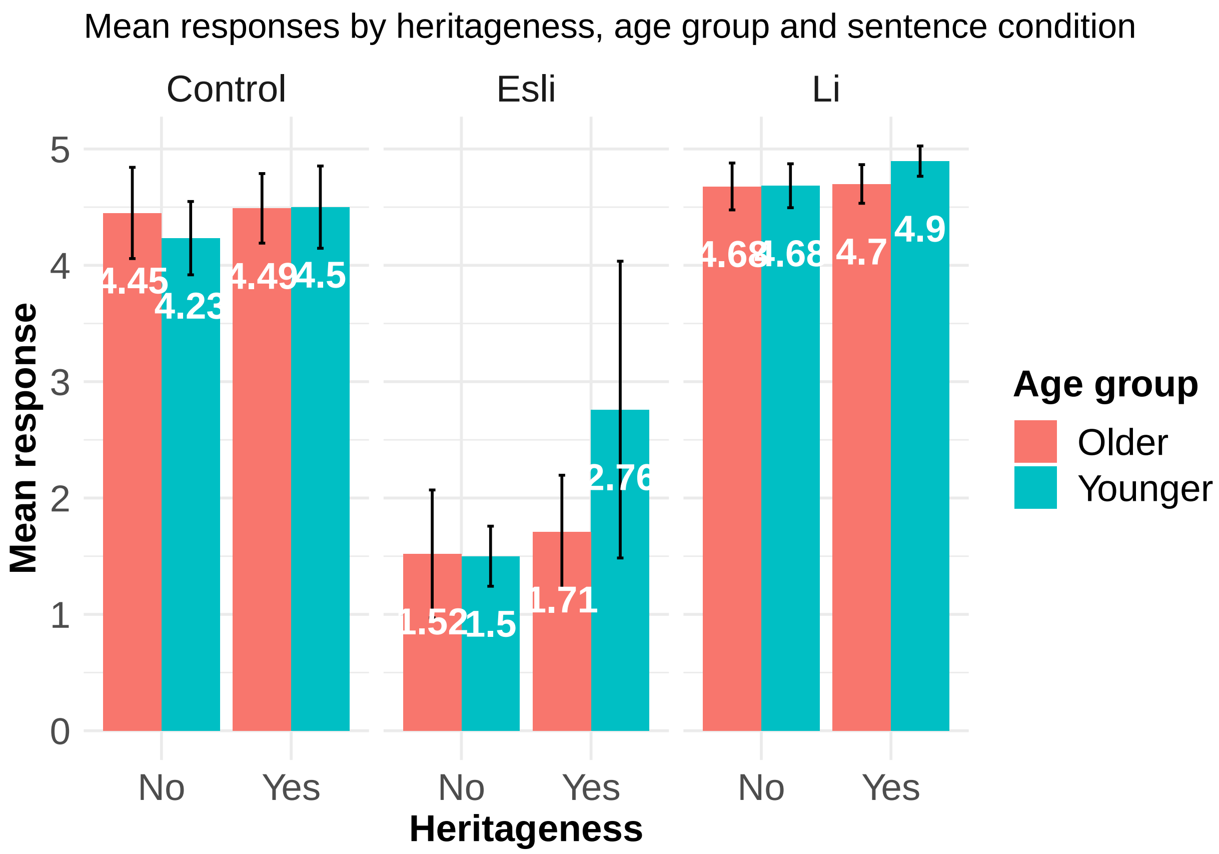
```
## # A tibble: 2 x 6
## # Groups:   heritageness [1]
##   heritageness age_group count_n mean_y_InUSA min_y_InUSA max_y_InUSA
##   <fct>        <fct>       <int>        <dbl>       <dbl>       <dbl>
## 1 yes          old            12         22            10          39
## 2 yes          young           5         11.6           4          22
```

Analysis

We will first check the data without droping the observations by RTs

Visualisation

```
# !!! We threat the ordinal vsriable (response) as numeric for the sake of visualisation. Need to think

facet_labels <- c(
                  `distr` = "Control",
                  `esli` = "Esli",
                  `li` = "Li"
                    )

working_data %>% group_by(subj_id, condition, age_group, heritageness) %>%
  summarise(mean_subj = mean(as.numeric(response))) %>%
  group_by(condition, age_group, heritageness) %>%
  summarise(mean_resp = mean(mean_subj), sd_resp = sd(mean_subj), se_resp = sd_resp/sqrt(n())) %>%
  ggplot(aes(heritageness, mean_resp, fill = age_group))+
  geom_bar(stat = "identity", position=position_dodge())+
  geom_errorbar(aes(ymin = mean_resp - 2*se_resp,
                    ymax = mean_resp + 2*se_resp), width = 0.1, position=position_dodge(0.9))+
  geom_text(aes(label=round(mean_resp,2)), vjust=3, color="white",
            position = position_dodge(0.9), size=5, fontface = "bold" )+
#  scale_y_continuous(breaks = seq(0,1,0.05))+
  ggtitle("Mean responses by heritageness, age group and sentence condition")+
  facet_grid(.~condition, labeller = as_labeller(facet_labels))+
  theme_minimal()+
  scale_fill_discrete(labels = c("Older", "Younger"))+
  scale_x_discrete(labels = c("No", "Yes"))+
  theme(axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"),
        strip.text.x = element_text(size = 14),
        legend.title = element_text(size = 14, face = "bold"),
        legend.text = element_text(size = 14)
        ) +
  labs(x = "Heritageness", y = "Mean response", fill = "Age group")
```



Ordinal regression

```
# check the response variable
working_data$response[1:10]
```

```
##  [1] "1" "5" "5" "2" "5" "5" "1" "4" "5" "1"
```

```
#now it is a factor variable, and we need an ordinal one


working_data$response_ordered <- ordered(working_data$response, levels = 1:5,
                              labels = c('absolutely_unacceptable',
                                         'mostly_unacceptable',
                                         'indefinite',
                                         'mostly_acceptable',
                                         'absolutely_acceptable'))


working_data$response_ordered[1:10]
```

```
##  [1] absolutely_unacceptable absolutely_acceptable
##  [3] absolutely_acceptable   mostly_unacceptable
##  [5] absolutely_acceptable   absolutely_acceptable
##  [7] absolutely_unacceptable mostly_acceptable
##  [9] absolutely_acceptable   absolutely_unacceptable
## 5 Levels: absolutely_unacceptable < ... < absolutely_acceptable
```

```
test_data <- working_data %>% select(response_ordered,
                                     condition,
                                     age_group,
                                     subj_id,
                                     coding,
                                     heritageness) %>% arrange(subj_id)



# I first tried this package but it took ages to compute
#install.packages("mixor")
#library(mixor)
#fit <- mixor(response ~ condition+age_group*heritageness,
#             data = test_data,
#             id = subj_id,
#             link = "logit")
#summary(fit)




#install.packages("ordinal")
library(ordinal)
```

```
## Warning: package 'ordinal' was built under R version 3.5.3
```

```
##
## Attaching package: 'ordinal'
```

```
## The following object is masked from 'package:dplyr':
##
##     slice
```

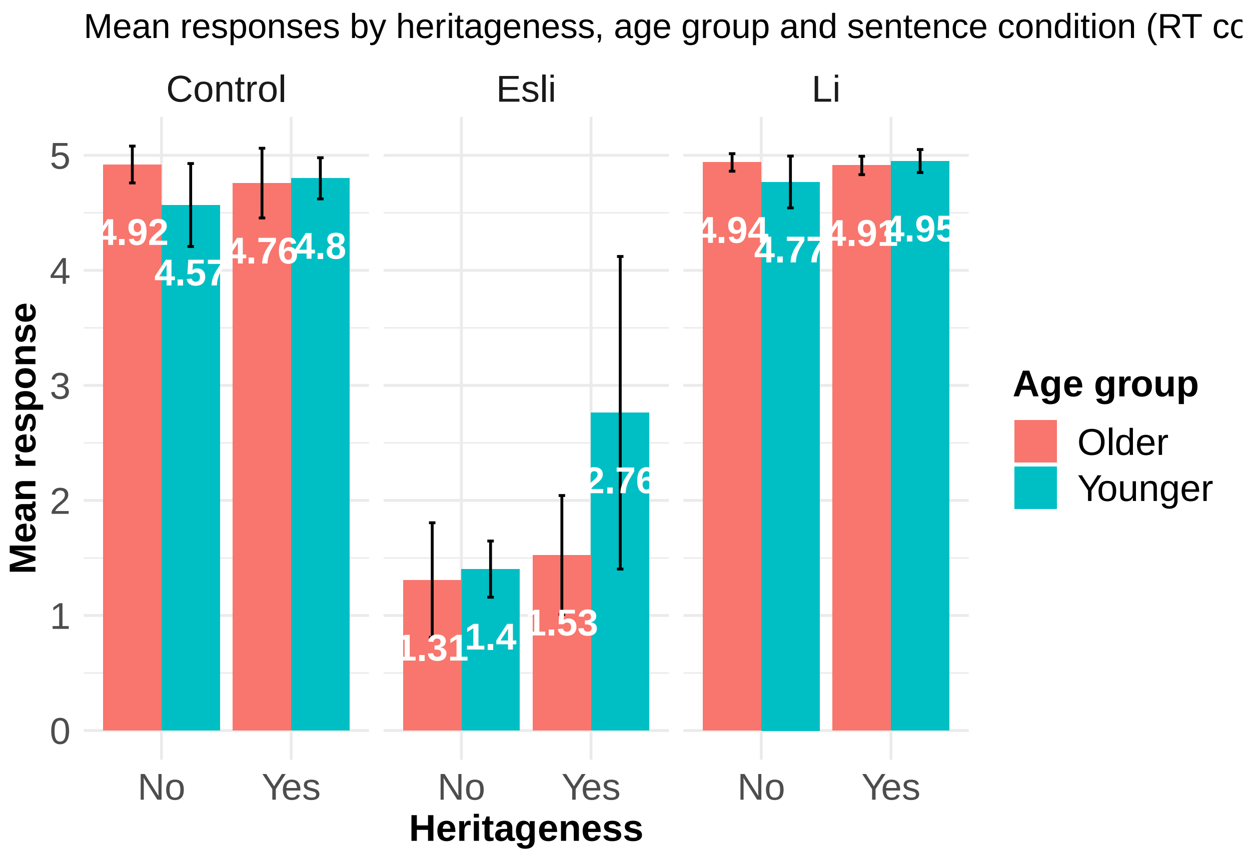
```
mod <-  clmm(response_ordered~condition*age_group*heritageness+(1|subj_id)+(1|coding), data=test_data,
             Hess=T
)


summary(mod)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: response_ordered ~ condition * age_group * heritageness + (1 |
##     subj_id) + (1 | coding)
## data:    test_data
```

```
##
##  link  threshold nobs logLik  AIC     niter      max.grad cond.H
##  logit flexible  1018 -916.12 1866.25 1614(4917) 1.54e-03 7.6e+02
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  coding  (Intercept) 0.3253   0.5703
##  subj_id (Intercept) 0.4936   0.7026
## Number of groups:  coding 47,  subj_id 37
##
## Coefficients:
##                                              Estimate Std. Error z value
## conditionesli                                 -5.4718     0.5643  -9.696
## conditionli                                    0.9016     0.5759   1.566
## age_groupyoung                                -0.3917     0.5379  -0.728
## heritagenessyes                                0.3168     0.5668   0.559
## conditionesli:age_groupyoung                   0.6143     0.5420   1.133
## conditionli:age_groupyoung                     0.2423     0.5953   0.407
## conditionesli:heritagenessyes                  0.3802     0.5709   0.666
## conditionli:heritagenessyes                   -0.4828     0.6291  -0.767
## age_groupyoung:heritagenessyes                 0.2293     0.7868   0.291
## conditionesli:age_groupyoung:heritagenessyes   1.4084     0.7656   1.839
## conditionli:age_groupyoung:heritagenessyes     0.7674     0.9175   0.836
##                                              Pr(>|z|)
## conditionesli                                  <2e-16 ***
## conditionli                                    0.1174
## age_groupyoung                                 0.4665
## heritagenessyes                                0.5761
## conditionesli:age_groupyoung                   0.2570
## conditionli:age_groupyoung                     0.6840
## conditionesli:heritagenessyes                  0.5054
## conditionli:heritagenessyes                    0.4429
## age_groupyoung:heritagenessyes                 0.7708
## conditionesli:age_groupyoung:heritagenessyes   0.0659 .
## conditionli:age_groupyoung:heritagenessyes     0.4029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                                             Estimate Std. Error z value
## absolutely_unacceptable|mostly_unacceptable  -4.5253     0.5441  -8.318
## mostly_unacceptable|indefinite               -2.9900     0.5274  -5.670
## indefinite|mostly_acceptable                 -2.0072     0.5177  -3.877
## mostly_acceptable|absolutely_acceptable      -0.8393     0.5114  -1.641
```

Now we will try to cut off the responses with RTs which more then 20% bigger then median (within each participant)

```r
# This is some trials to take into account the rt data

data_adjust_rt <- working_data %>%
  group_by(subj_id) %>%
  mutate(reference = median(response_time)*1.2) %>%
  filter(response_time<=reference)

data_adjust_rt %>% group_by(subj_id, condition, age_group, heritageness) %>%
  summarise(mean_subj = mean(as.numeric(response))) %>%
  group_by(condition, age_group, heritageness) %>%
  summarise(mean_resp = mean(mean_subj), sd_resp = sd(mean_subj), se_resp = sd_resp/sqrt(n())) %>%
  ggplot(aes(heritageness, mean_resp, fill = age_group))+
  geom_bar(stat = "identity", position=position_dodge())+
  geom_errorbar(aes(ymin = mean_resp - 2*se_resp,
                    ymax = mean_resp + 2*se_resp), width = 0.1, position=position_dodge(0.9))+
  geom_text(aes(label=round(mean_resp,2)), vjust=3, color="white",
            position = position_dodge(0.9), size=5, fontface = "bold" )+
#  scale_y_continuous(breaks = seq(0,1,0.05))+
  ggtitle("Mean responses by heritageness, age group and sentence condition (RT corrected data)")+
  facet_grid(.~condition, labeller = as_labeller(facet_labels))+
  theme_minimal()+
  scale_fill_discrete(labels = c("Older", "Younger"))+
  scale_x_discrete(labels = c("No", "Yes"))+
  theme(axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"),
        strip.text.x = element_text(size = 14),
        legend.title = element_text(size = 14, face = "bold"),
        legend.text = element_text(size = 14)
        ) +
  labs(x = "Heritageness", y = "Mean response", fill = "Age group")
```



```
# check the response variable
data_adjust_rt$response[1:10]
```

```
##  [1] "5" "5" "5" "5" "1" "4" "1" "5" "1" "5"
```

```
#now it is a factor variable, and we need an ordinal one


data_adjust_rt$response_ordered <- ordered(data_adjust_rt$response, levels = 1:5,
                              labels = c('absolutely_unacceptable',
                                         'mostly_unacceptable',
                                         'indefinite',
                                         'mostly_acceptable',
                                         'absolutely_acceptable'))


data_adjust_rt$response_ordered[1:10]
```

```
##  [1] absolutely_acceptable   absolutely_acceptable
##  [3] absolutely_acceptable   absolutely_acceptable
##  [5] absolutely_unacceptable mostly_acceptable
##  [7] absolutely_unacceptable absolutely_acceptable
##  [9] absolutely_unacceptable absolutely_acceptable
## 5 Levels: absolutely_unacceptable < ... < absolutely_acceptable
```

```
test_data <- data_adjust_rt %>% select(response_ordered,
                                     condition,
                                     age_group,
                                     subj_id,
                                     coding,
                                     heritageness) %>%
  arrange(subj_id)

mod_rt <-  clmm(response_ordered~condition*age_group*heritageness+
                  (1|subj_id)+
                  (1|coding), data=test_data,
             Hess=T
)


summary(mod_rt)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: response_ordered ~ condition * age_group * heritageness + (1 |
##     subj_id) + (1 | coding)
## data:    test_data
##
##  link  threshold nobs logLik  AIC    niter      max.grad cond.H
##  logit flexible  615  -385.18 804.36 1485(5870) 3.12e-04 1.2e+03
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  coding  (Intercept) 7.302e-10 2.702e-05
##  subj_id (Intercept) 1.489e+00 1.220e+00
## Number of groups:  coding 47,  subj_id 37
##
## Coefficients:
##                                              Estimate Std. Error z value
## conditionesli                                 -8.4406     1.0335  -8.167
## conditionli                                    0.2583     1.0918   0.237
## age_groupyoung                                -1.2005     1.0747  -1.117
## heritagenessyes                               -0.3673     1.1228  -0.327
## conditionesli:age_groupyoung                   1.7419     1.0614   1.641
## conditionli:age_groupyoung                     0.5797     1.2032   0.482
## conditionesli:heritagenessyes                  1.4026     1.1059   1.268
## conditionli:heritagenessyes                    0.1080     1.2749   0.085
## age_groupyoung:heritagenessyes                 0.4951     1.4650   0.338
## conditionesli:age_groupyoung:heritagenessyes   1.6471     1.4036   1.173
## conditionli:age_groupyoung:heritagenessyes     0.2501     1.6658   0.150
##                                              Pr(>|z|)
## conditionesli                                3.16e-16 ***
## conditionli                                     0.813
## age_groupyoung                                  0.264
## heritagenessyes                                 0.744
## conditionesli:age_groupyoung                    0.101
## conditionli:age_groupyoung                      0.630
## conditionesli:heritagenessyes                   0.205
## conditionli:heritagenessyes                     0.933
## age_groupyoung:heritagenessyes                  0.735
## conditionesli:age_groupyoung:heritagenessyes    0.241
## conditionli:age_groupyoung:heritagenessyes      0.881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                                             Estimate Std. Error z value
## absolutely_unacceptable|mostly_unacceptable  -6.7906     1.0290  -6.599
## mostly_unacceptable|indefinite               -4.8370     0.9950  -4.861
## indefinite|mostly_acceptable                 -3.9944     0.9791  -4.080
## mostly_acceptable|absolutely_acceptable      -2.8767     0.9618  -2.991
```

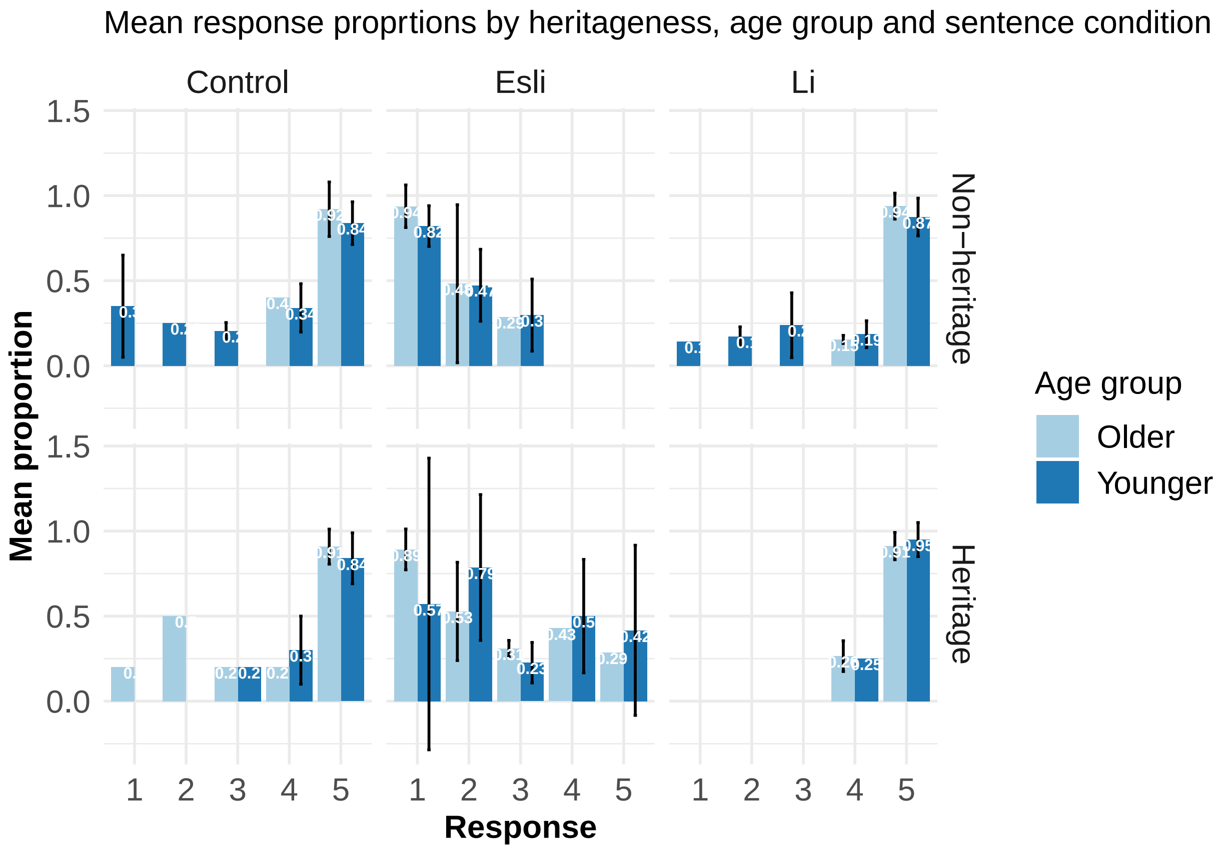
```
summary_mod <- summary(mod_rt)
```

```
data_summary <- data_adjust_rt %>% mutate(response=as.factor(response)) %>%
  group_by(subj_id, condition, age_group, heritageness) %>%
  mutate(all = n()) %>%
  group_by(subj_id, condition, age_group, heritageness, response, all) %>%
  summarise(quant = n()) %>%
  mutate(prop = quant/all) %>%
  ungroup() %>%
  group_by(condition, age_group, heritageness, response) %>%
  summarise(mean_prop = mean(prop), sd_prop = sd(prop), se_prop = sd_prop/sqrt(n()))

data_summary
```

```
## # A tibble: 42 x 7
## # Groups:   condition, age_group, heritageness [12]
##    condition age_group heritageness response mean_prop  sd_prop  se_prop
##    <fct>     <fct>     <fct>        <chr>        <dbl>    <dbl>    <dbl>
##  1 distr     old       no           4            0.4    NaN      NaN
##  2 distr     old       no           5            0.92     0.179    0.08
##  3 distr     old       yes          1            0.2    NaN      NaN
##  4 distr     old       yes          2            0.5    NaN      NaN
##  5 distr     old       yes          3            0.2    NaN      NaN
##  6 distr     old       yes          4            0.2    NaN      NaN
##  7 distr     old       yes          5            0.908    0.178    0.0514
##  8 distr     young     no           1            0.35     0.212    0.15
##  9 distr     young     no           2            0.25   NaN      NaN
## 10 distr     young     no           3            0.206    0.0419   0.0242
## # ... with 32 more rows
```

```
facet_labels <- c(
  `distr` = "Control",
  `esli` = "Esli",
  `li` = "Li",
  `yes` = "Heritage",
  `no` = "Non-heritage"
)

data_summary %>%
  ggplot(aes(response, mean_prop, fill = age_group))+
  geom_bar(stat = "identity", position=position_dodge(0.9, preserve = "single"))+
  geom_errorbar(aes(ymin = mean_prop - 2*se_prop,
                    ymax = mean_prop + 2*se_prop), width = 0.13,
                position=position_dodge(0.9, preserve = "single"), color = "black")+
  facet_grid(heritageness~condition, labeller = as_labeller(facet_labels)) +
  geom_text(aes(label=round(mean_prop,2)), vjust=1, color="white",
            position = position_dodge(width = 0.9), size=2, fontface = "bold")+
  ggtitle("Mean response proprtions by heritageness, age group and sentence condition")+
  theme_minimal()+
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        strip.text.x = element_text(size = 12),
        strip.text.y = element_text(size = 12),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 12),
        plot.title = element_text(size = 12)) +
  labs(x = "Response", y = "Mean proportion", fill = "Age group")+
  scale_fill_brewer(labels = c("Older", "Younger"), palette = "Paired")
```

```
## Warning: Removed 12 rows containing missing values (geom_errorbar).
```


Mean response proprtions by heritageness, age group and sentence condition
Control
Esli
Li
Non−heritage
Heritage
Mean proportion
Response
Age group
Older
Younger