Preliminary Data Screening

Kaleem Ullah

2024-03-15

Check the percentage of videos that were played

```
# setwd('D:\\RA - Spring 2024\\Gorilla Data\\raw_data')
video_data_recog <- read.csv("metadata/recog_flat.csv", header=TRUE)
video_data_iden <- read.csv("metadata/iden_flat.csv", header=TRUE)
video_data_recog['trial'] = video_data_recog['trial.number']
video_data_iden['trial'] = video_data_iden['trial.number']

num_ppn_recog <- length(unique(video_data_recog$ppn))
num_ppn_iden <- length(unique(video_data_iden$ppn))
df_recog <- read.csv("raw_data/emotion_recog.csv", header=TRUE)
df_iden <- read.csv("raw_data/emotion_iden.csv", header=TRUE)
expected_played_recog <- num_ppn_recog*24
expected_played_iden <- num_ppn_iden*24
videos_played_recog <- length(video_data_recog$reaction.recog)
videos_played_iden <- length(video_data_iden$reaction.iden)
```

```
head(video_data_recog)
```

```
##                 stimuli trial.number      ppn reaction.recog trial
## 1 AD_pp01_pos_anger.mp4            3 10607570        1290.70     3
## 2 AD_pp01_pos_anger.mp4            4 10813367        1007.30     4
## 3 AD_pp01_pos_anger.mp4            6 10750001         925.00     6
## 4 AD_pp01_pos_anger.mp4            7 10603446         512.30     7
## 5 AD_pp01_pos_anger.mp4            7 10749149         173.38     7
## 6 AD_pp01_pos_anger.mp4           13 10781304       14798.00    13
```

Check individually for each task

Expected videos played: 3.24×10^4

Total videos played: 30588

Percentage of videos played: 94.4074074

1. Recognition Task

Videos Played

```
zoneType <- df_recog$Zone.Type
last_iter = "response_button_text"
unplayed = 0
for (i in zoneType){
  if (last_iter == "response_button_text" & i  == "response_button_text") {
    unplayed = unplayed + 1
```

```
  }
  last_iter = i
}
unplayed / length(df_recog$Zone.Type)
```

```
## [1] 0.02517328
```

Reaction Time

```
reac_recog <- sum(video_data_recog$reaction.recog < 1000, na.rm=TRUE)
```

Percentage of videos played less than 1 second: 50.8794298

Time Taken for each task

```
total_time_recog <- as.numeric(df_recog[df_recog$Trial.Number == "END TASK",]$Reaction.Time) / 1000
```

```
summary(total_time_recog)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     2.667   168.924   228.696   322.870   337.230 12100.291
```

2. Identification Task

Expected videos played: 3.2208×10^4

Total videos played: 30363

Percentage of videos played: 94.2716095

```
reac_iden <- sum(video_data_iden$reaction.iden < 1000, na.rm=TRUE)
```

Percentage of videos played less than 1 second: 48.3228717

```
total_time_iden <- as.numeric(df_iden[df_iden$Trial.Number == "END TASK",]$Reaction.Time) / 1000
```

```
summary(total_time_iden)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    3.427  138.265  194.745  273.549  274.475 5333.155
```

Percentage of videos watched by participants

Recognition

Check number of videos played

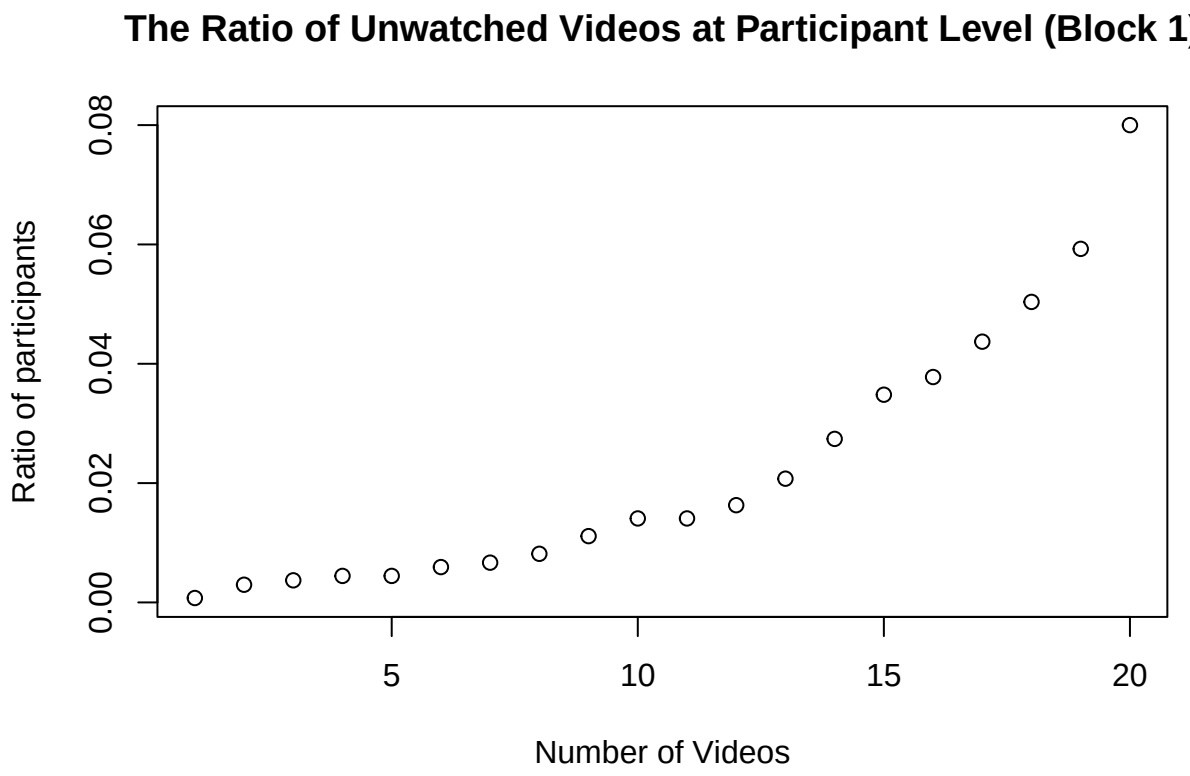
```
num_ppn <- length(unique(video_data_recog$ppn))
number_per_ppn <- numeric(num_ppn)
ppns <- as.list(unique(video_data_recog['ppn']))
i = 0
for (each_ppn in ppns$ppn) {
  # print(each_ppn)
  # print()
  number_per_ppn[i] <- length(video_data_recog$trial[video_data_recog["ppn"] == each_ppn])
  # print(numbers_per_ppn)
  i = i + 1
}
```

Create csv with number of videos watched

```
videos_per_ppn <- data.frame(ppn = ppns, videos = number_per_ppn)
write.csv(videos_per_ppn, 'metadata/videos_per_ppn_1.csv')
```

```
ratio_vids <- numeric(23)
for (i in 1:24) {
  # print(i)
  ratio_vids[i] <- sum(number_per_ppn < i) / num_ppn
}
```

```
# png("plots/ratio_recog.png", width=800, height=600)
plot(ratio_vids[1:20], main = "The Ratio of Unwatched Videos at Participant Level (Block 1)", ylab="Rati
```




```
# dev.off()
```

```
print(sum(number_per_ppn < 12) / num_ppn)
```

```
## [1] 0.0162963
```

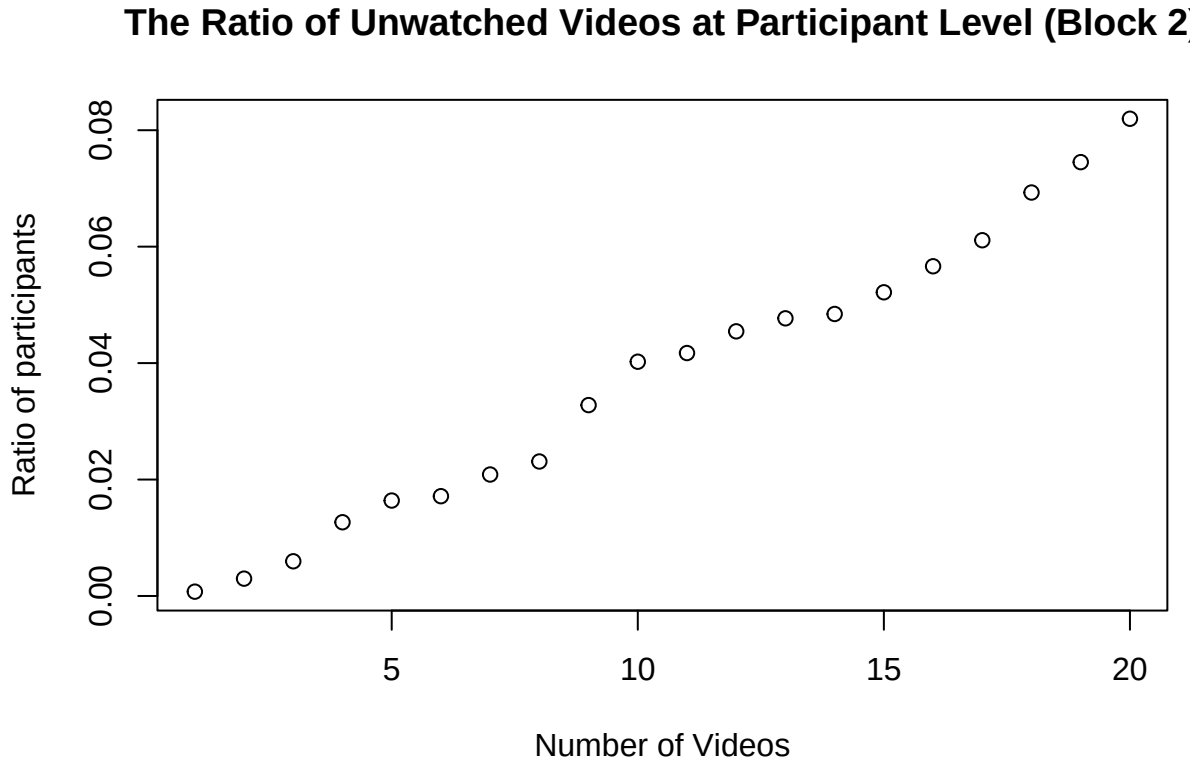
Identification

```
num_ppn <- length(unique(video_data_iden$ppn))
number_per_ppn <- numeric(num_ppn)
ppns <- as.list(unique(video_data_iden['ppn']))
i = 0
for (each_ppn in ppns$ppn) {
```

```
  # print(each_ppn)
  # print()
  number_per_ppn[i] <- length(video_data_iden$trial[video_data_iden["ppn"] == each_ppn])
  # print(numbers_per_ppn)
  i = i + 1
}
```

```
ratio_vids <- numeric(23)
for (i in 1:24) {
  # print(i)
  ratio_vids[i] <- sum(number_per_ppn < i) / num_ppn
}
```

```
# png("plots/ratio_iden.png", width=800, height=600)
plot(ratio_vids[1:20], main = "The Ratio of Unwatched Videos at Participant Level (Block 2)", ylab="Rati
```




```
# dev.off()
```

```
print(sum(number_per_ppn < 12) / num_ppn)
```

```
## [1] 0.04545455
```

```
for (i in 1:24) {
  # print(i)
  print(sum(number_per_ppn < i) / length(number_per_ppn))
}
```

```
## [1] 0.0007451565
```

```
## [1] 0.002980626
## [1] 0.005961252
## [1] 0.01266766
## [1] 0.01639344
## [1] 0.0171386
## [1] 0.02086438
## [1] 0.02309985
## [1] 0.03278689
## [1] 0.04023845
## [1] 0.04172876
## [1] 0.04545455
## [1] 0.04769001
## [1] 0.04843517
## [1] 0.05216095
## [1] 0.05663189
## [1] 0.06110283
## [1] 0.06929955
## [1] 0.07451565
## [1] 0.08196721
## [1] 0.09090909
## [1] 0.1177347
## [1] 0.1743666
## [1] 0.2578241
```

Accuracy Check in difference for reaction time

```
less_than_one <- video_data_recog[video_data_recog$reaction.recog < 1000, ]
more_than_one <- video_data_recog[video_data_recog$reaction.recog > 1000, ]
```

```
clean_data <- read.csv("new_csv/preprocessed_data.csv", header = TRUE)
```

```
join_less <- clean_data %>% right_join(less_than_one, by=c("trial", 'ppn'), relationship = "many-to-many
join_less <- na.omit(join_less)
sum(join_less$accuracy) / length(join_less$accuracy)
```

```
## [1] 0.1889255
```

```
join_more <- clean_data %>% right_join(more_than_one, by=c("trial", 'ppn'), relationship = "many-to-many
join_more <- na.omit(join_more)
sum(join_more$accuracy) / length(join_more$accuracy)
```

```
## [1] 0.1754538
```