W241 Final Project - Gender and Following Directions

Daniel Alvarez, Austin Doolittle, and Bethany Keller

4/18/2020

> "I personally take the subway in New York. I love the subway because of the male and female voices that narrate the subway... that man and that woman. The woman who does the subway announcements and her way too loud husband. You know the woman who is like, (in a soft female voice) "The next stop is Christopher Street." (in a strong male voice) "STAND CLEAR OF THE CLOSING DOORS PLEASE!". I asked my friend who works for the city, why is it a male voice and a female voice. And he told me because it's been proven that people take information from a female voice, but they will only take a warning from a male voice. Now that's its own American gender nightmare that we don't have time to get into..." — Mulaney (2019)

Introduction

Implicit bias surrounding gender is ever-present in society and, in recent times, more attention has been brought to the issue. The increased attention has given rise to more effort to identify gender biases and attempt to mitigate and remove them. However, even with increased attention to the issue, gender biases persist in our everyday life - from the technologies we use such as voice assistants on smartphones and the audio commands in public service announcements. A CNN article (October 2011) discusses how the choice of the gendered voice of the Siri audio assistant in Apple I-Phone smartphone, which is female in the US market, has to do with biological studies that suggest that people generally find women's voices more pleasing than men's. The same article cites historical references such as the use of female voices in navigation devices during World War II, when women's voices were employed in airplane cockpits because they stood out among the male pilots. Moreover, "telephone operators have traditionally been female, making people accustomed to getting assistance from a disembodied woman's voice." (Griggs 2011) Recent media reports also cite cases of women in high positions deepening their voices to sound more baritone in order to convey authority (for example, the case of Theranos CEO Elizabeth Holmes) (Hesse 2019). With this motivation, we pursued identifying the existence of gender bias in giving directions through an experimental study.

Research Question

The experimental study attempts to isolate the effects that a speaker's gender has on the willingness of a targeted listener to follow through on the direction of the speaker. The purpose is to provide insight on whether the verbal directions of men are more effective than that of women in steering behavior though some simple game playing. The experiment should serve to examine the existence of the effect and the strength of the effect, if it exists.

Experimental Design

Discussion of the Treatment

In order to estimate the effect of a gender bias in following directions, we designed an online tic-tac-toe game whereby participants hear audio directive cues to make play moves in gendered voices (male or female). The treatment is the audio directive to make a given move while playing tic-tac-toe against a computer adversary as well as a visual indicator over the suggested square to make a move. The control group receives no audio directive, but rather, a visual indicator over the suggested square to make a move. Audio (for the treatment) and visual (for the control) suggestions are provided randomly with some moves provided as optimal moves and other moves as random, likely sub-optimal, suggestions following a random process.

There are two measured outcomes in the study: (1) the proportion p_{all} of all moves that the subject follows the recommendation of the directive over the course of playing the online tic-tac-toe games; (2) the proportion $p_{optimal}$ of optimal moves that the subject follows the recommendation of the directive over the course of playing the online tic-tac-toe games. The measured outcomes in (1) would presumably prevent the subject from always taking the suggested directive through natural game play since the subjects may receive random sub-optimal directives. If all suggested moves are optimal, as in the case of measured outcomes in (2), the proportion of moves followed would indicate how well subjects follow suggested directives given the treatment received.

The treatment effect is the mean difference in outcomes between the treatment and control group, given the treatment is received. The experimental concern of interest is how the gender of the speaker affects the experimental subject's willingness to listen in direct comparison to the control group's willingness to listen. Using the measured outcomes in (1) attempts to capture the generalized treatment effect for all suggested moves. Using the measured outcomes in (2) attempts to capture the treatment effect for optimal suggested moves.

Comparison of Potential Outcomes

The potential outcomes are the proportion of moves made by an experimental subject that follow the directive suggestions in the course of tic-tac-toe game play. There are three types of potential outcomes in the experiment corresponding to the treatment assignments: control (no audio directive), treatment-female (female audio voice directive), and treatment-male (male audio voice directive). Those assigned to receive the control will see a visual suggestion with a square highlighted in the tic-tac-toe board. Formally, the potential outcomes in this study can be described as:

$Y_i(TM = 1)$: random variable representing the potential outcomes for a given experimental i^{th} subject receiving the male voice audio treatment. $Y_i(TF = 1)$: random variable representing the potential outcomes for a given experimental i^{th} subject receiving the female voice audio treatment. $Y_i(T = 0)$: random variable representing the potential outcomes for a given experimental i^{th} subject receiving the control of no audio treatment, but rather a square highlighted in the tic-tac-toe board.

We can be assured through the construction of the online tic-tac-toe game and careful assignment process that those subjects assigned to the treatment groups will receive the treatment. In other words, those assigned to the treatment group, will not inadvertently not receive the control. Therefore, we can state that the expected potential outcome for given subject randomly allocated to a treatment group is equal to the expected outcome for a given subject randomly allocated to a treatment group conditional on the subject actually receiving the treatment; formally, $E[Y_i(1)] = E[Y_i(1)|d_i = 1]$ for both the treatment-male and treatment-female voice recipients.

The average treatment effect (ATE) can be described as the difference in average outcomes in the treatment and control groups. Formally, this is written as:

The difference in average potential outcomes for subjects in the treatment-male group from the average potential outcomes for subjects in the control group:

$E[Y_i(TM = 1)|D_i = 1] - E[Y_i(T = 0)|D_i = 0]$

The difference in average potential outcomes for subjects in the treatment-female group from the average potential outcomes for subjects in the control group:

$E[Y_i(TF = 1)|D_i = 1] - E[Y_i(T = 0)|D_i = 0]$

Subjects are randomly assigned to treatment and control groups and therefore, there is no selection bias and we can say the ATE is unbiased. The ATE among randomly treated subjects is the same as the ATE among all subjects.

We also compare the difference in average potential outcomes across the two treatment groups for further exposition; although this will be lower-powered as we discuss below.

Random Assignment

Randomization is done by taking all experimental subjects and assigning roughly half of participants to the control group and the other half to the treatment group, using the `sample` function in R. Given our suspicion that the gender of the experimental subject may influence how (s)he may respond to the treatment, a blocking design was used on the subject's gender. In this blocking design, a roughly equal proportion of male subjects and female subjects were assigned randomly to treatment and control groups. Subjects in the treatment group were then further randomly assigned to treatment-male and treatment-female groups.

The following R code demonstrates the randomization approach and subject gender-blocking approach:

Set an `id` variable (ordinal) useful for randomization assignment, create a binary numerical gender variable, and rename the long variable names from the pre-treatment survey data.

```
d[,'id'] <- c(1:nrow(d))
colnames(d)[1:5] <-c("timestamp","gender_string","age","email","consent")
d[, gender:=ifelse(gender_string=='Female',0,1)]
dim(d)
```

Importantly, we check for any "No" consents. These would be excluded from the study.

```
table(d$consent)
```

Examine the gender and age distibution.

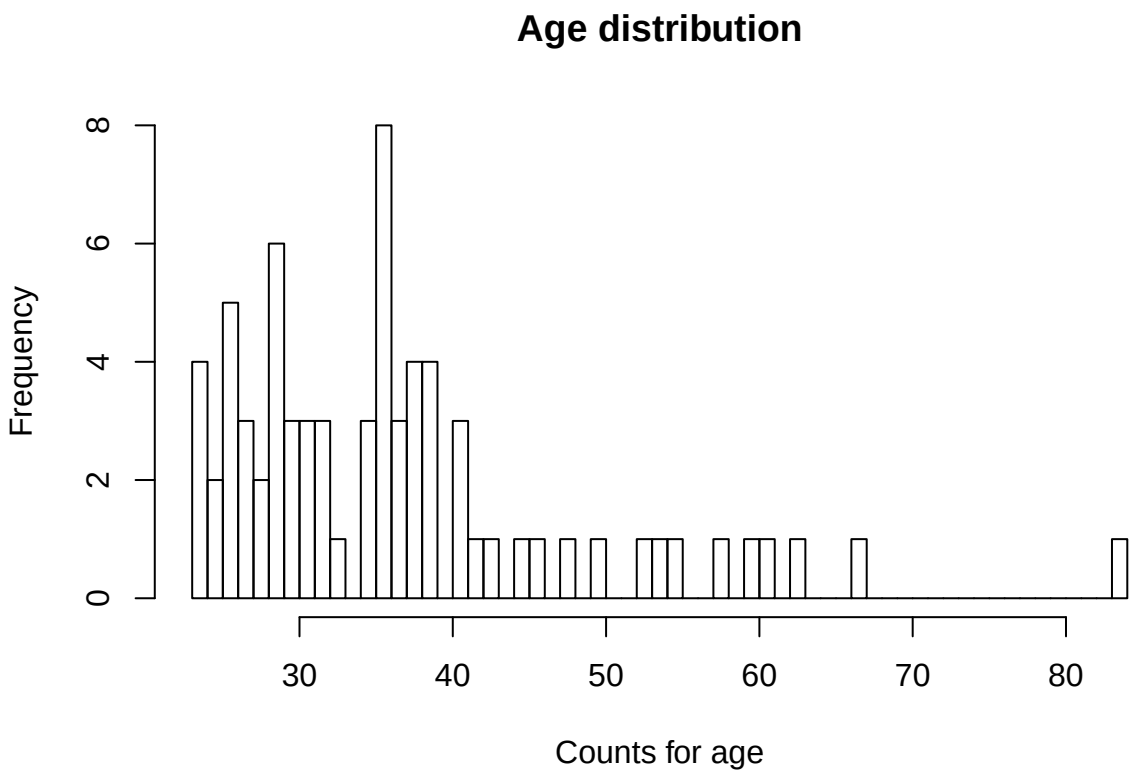
```
#crosstab of gender
table(d$gender_string)
```

```
##
## Female   Male
##     32     40
```

```
# summary statistics of age
summary(d$age, na.omit=FALSE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   29.00   36.00   36.96   39.50   84.00
```

```
# histogram of age
hist(d$age,main="Age distribution", xlab="Counts for age",ylab = "Frequency", breaks=50)
```



We find there are more men then women in the study (32 females and 40 males). The age distribution is right-skewed with most subjects tending to be in their 30s (median of 36).

Perform a block randomization on gender to assign treatment in approximately equal proportions to male and female subjects.

```
# Assign count variables
N = nrow(d)
n_female <- length(d$age[d$gender==0])
n_male = length(d$age[d$gender==1])

#randomly assign subjects to either treatment or control groups blocking by state
randomize_blockbygender <- function(){
  ifelse(d$gender==1,sample(c(rep(0,n_female/2),rep(1,n_female/2))),
         sample(c(rep(0,n_male/2),rep(1,n_male/2))))
}

# create treatment assignment variable
d[,'treatment'] <- randomize_blockbygender()

#Check cross-tab of treatment and gender
kable(d[,xtabs(~gender+treatment)])
```

	0	1
0	18	14
1	19	21

```
#kable(table(d$gender,d$treatment))
```

Split treatment group further into two treatment groups for female `TF` and male voice `TM`.

```r
# assign treatment count variable
n_treatment = length(d$age[d$treatment==1])

#randomly assign subjects to either two treatment groups (TF and TM)
randomize_treatment <- function(){
  ifelse(d$treatment==1,sample(c(rep(0,n_treatment/2),rep(1,n_treatment/2))),0)
}

d[,'is_treatedmalevoice'] <- randomize_treatment()

#Check cross-tab of treatment and is_treatedmalevoice
kable(table(d$is_treatedmalevoice,d$treatment))
```

	0	1
0	37	17
1	0	18

For dataset readability purposes, include a string variable for each subjects assignment: control `C`, treatment with male voice `TM`, and treatment with female voice `TF`.

```r
# create a string assignment variable for readability purposes
d[treatment==0, assignment:='C']
d[treatment==1 & is_treatedmalevoice==1, assignment:='TM']
d[treatment==1 & is_treatedmalevoice==0, assignment:='TF']

# check cross-tab
table(d$assignment)
```

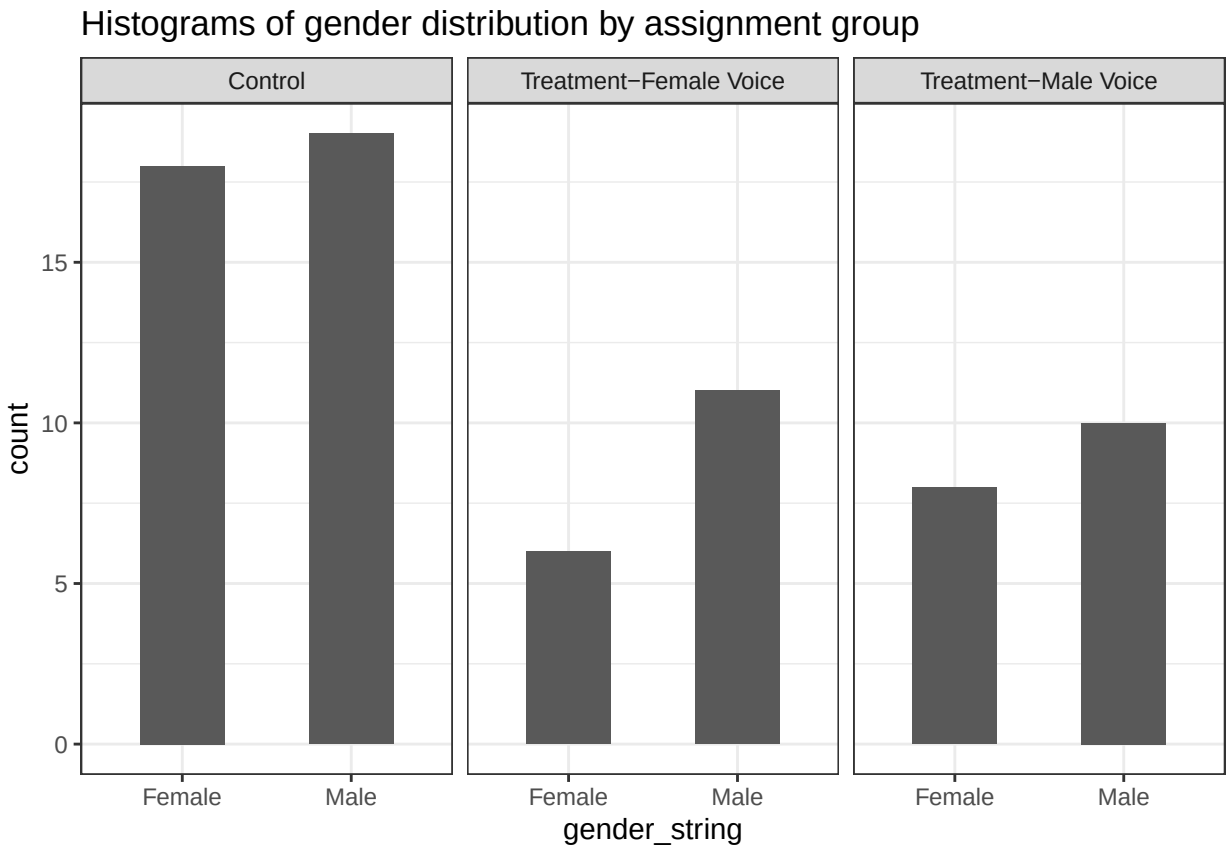
```
##
##  C TF TM
## 37 17 18
```

Following the assignment randomization process, we examine the balance of the subject gender across the treatment and control groups. We find an appropriate level of gender balance across groups that should improve the power of our treatment effect estimates. The higher count of male subjects across all treatment assignment groups reflects the fact that there are more male subjects in the study overall; however, the gender balance is proportionally consistent across treatment assignment groups.

```r
# check cross-tab of assignment and subject gender
table(d$assignment, d$gender_string)
```

```
##
##      Female Male
##   C      18   19
##   TF      6   11
##   TM      8   10
```

```r
assignment.labs <- c("Control", "Treatment-Female Voice", "Treatment-Male Voice")
names(assignment.labs) <- c("C", "TF", "TM")

ggplot(d,aes(x=gender_string))+stat_count(width = 0.5)+
  facet_grid(~assignment, labeller = labeller(assignment = assignment.labs))+
  labs(title = 'Histograms of gender distribution by assignment group')+
  theme_bw()
```

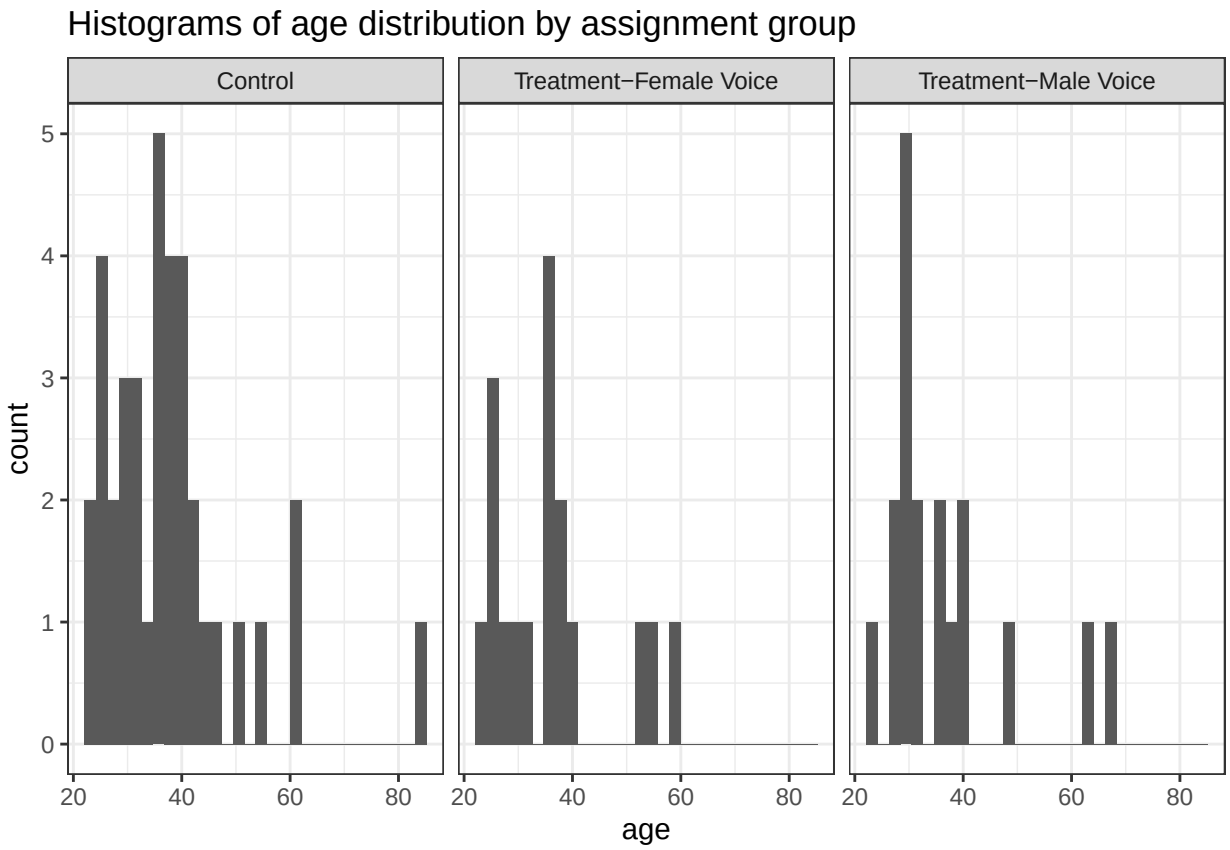


Although we did not perform block randomization on age, we also examine the balance on age and observe an approximately similar balance, although the group assigned the female audio treatment skewed slightly younger. We did not think this had a meaningful impact on results.

```
# check cross-tab of assignment and age
table(d$assignment, d$age)
```

```
##
##      23 24 25 26 27 28 29 30 31 32 33 35 36 37 38 39 41 42 43 45 46 48 50 53 54
##   C   0  2  0  4  1  1  3  0  2  1  1  0  5  2  2  2  2  1  1  1  1  0  1  0  1
##   TF  1  0  2  1  1  0  1  0  0  1  0  2  2  0  2  1  0  0  0  0  0  0  0  1  0
##   TM  1  0  0  0  1  1  2  3  1  1  0  1  1  1  0  1  1  0  0  0  0  1  0  0  0
##
##      55 58 60 61 63 67 84
##   C   0  0  1  1  0  0  1
##   TF  1  1  0  0  0  0  0
##   TM  0  0  0  0  1  1  0
```

```
# examine age distributions across assignment groups

assignment.labs <- c("Control", "Treatment-Female Voice", "Treatment-Male Voice")
names(assignment.labs) <- c("C", "TF", "TM")

ggplot(d,aes(x=age))+geom_histogram(bins=30)+
  facet_grid(~assignment, labeller = labeller(assignment = assignment.labs))+
  labs(title = 'Histograms of age distribution by assignment group')+
  theme_bw()
```



Discussion on Attrition

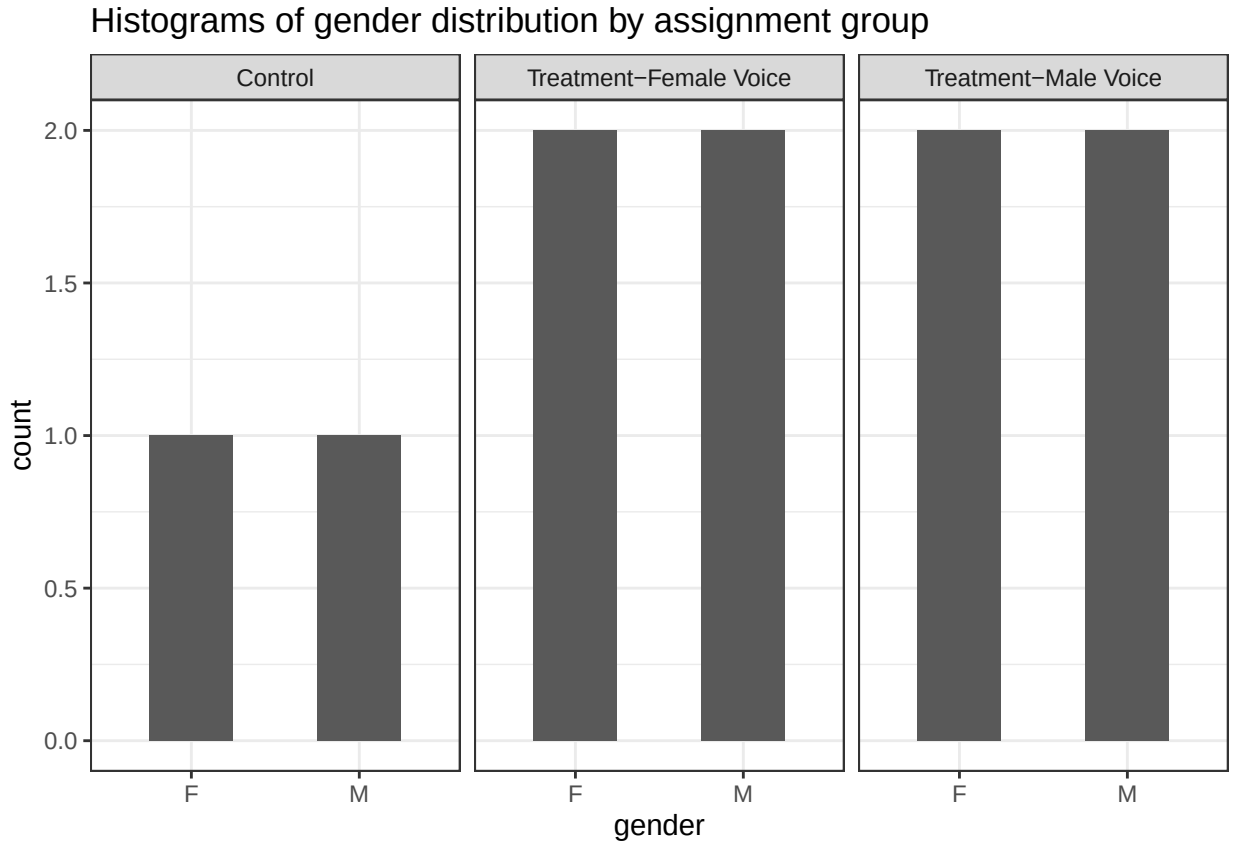
There were 10 subjects that completed the pre-experiment survey, yet did not actually take the treatment by playing the games. As a result, we observed missing outcome data for these 10 subjects. We viewed this as an innocuous form of attrition since the missing data for the attriters is independent of potential outcomes (missing independent of potential outcomes, $Y_i(z) \perp\!\!\!\perp R_i(z)$. This is satisfied by the research design since subjects were divided into random subgroups and outcomes were measured through different randomly assigned exposures unknown in advance. We observe that the attriters were approximately balanced across treatment assignment groups (2 in the control group, 4 in the male voice treatment and 4 in the female voice treatment) and therefore, we do not think that attrition is systematically related to a subject's potential outcomes. Moreover, in performing a covariate balance check for these attriters, we find these subjects approximately balanced across covariates. See the plot of the gender and age distribution for attriters below. Give the benign nature of this attrition, we choose to exclude these subjects with missing outcomes in analyzing the experimental results.

```
# check cross-tab of assignment and subject gender
kable(table(final_attriters$assignment, final_attriters$gender))
```

	F	M
C	1	1
TF	2	2
TM	2	2

```
assignment_status.labs <- c("Control", "Treatment-Female Voice", "Treatment-Male Voice")
names(assignment_status.labs) <- c("C", "TF", "TM")

ggplot(final_attriters,aes(x=gender))+stat_count(width = 0.5)+
```

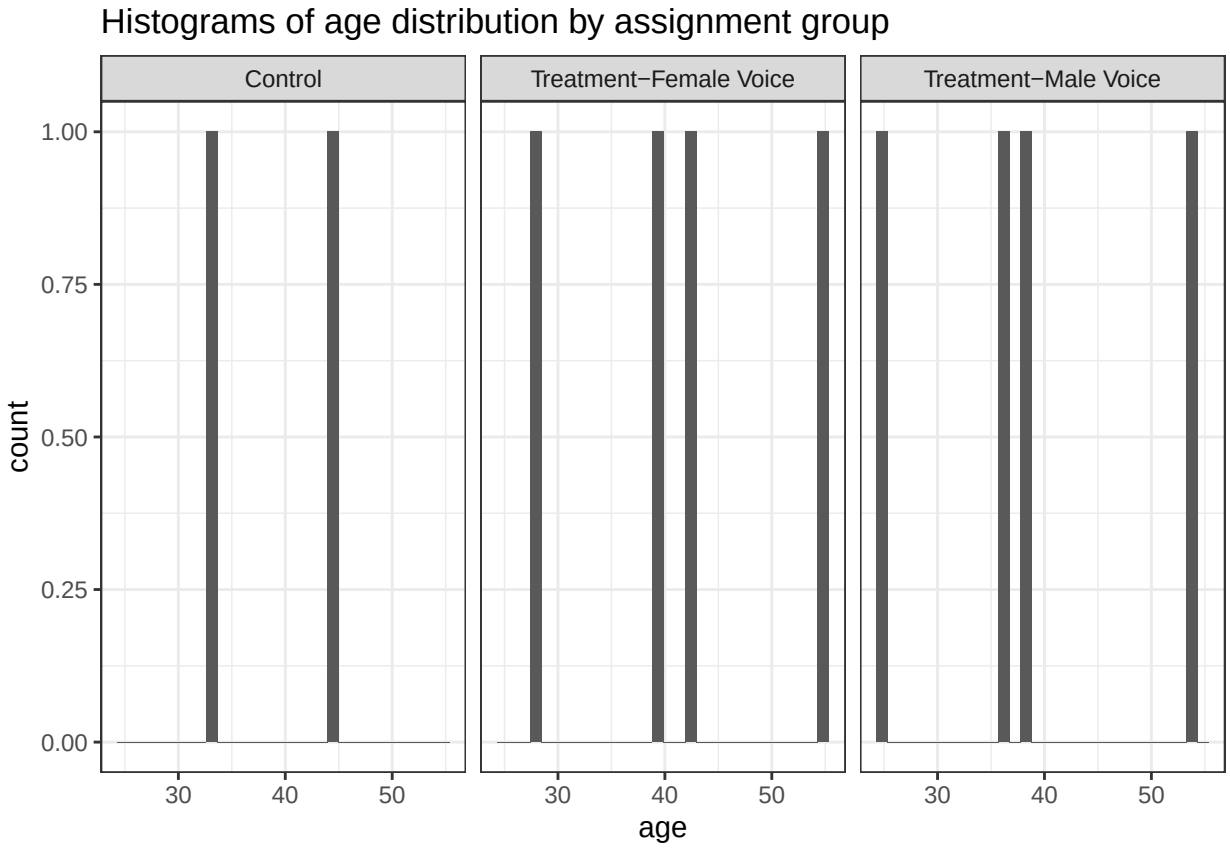
```
  facet_grid(~assignment_status, labeller = labeller(assignment_status = assignment_status.labs))+
  labs(title = 'Histograms of gender distribution by assignment group')+
  theme_bw()
```




```
# check cross-tab of assignment and age
kable(table(final_attriters$assignment, final_attriters$age))
```

	25	28	33	36	38	39	42	45	54	55
C	0	0	1	0	0	0	0	1	0	0
TF	0	1	0	0	0	1	1	0	0	1
TM	1	0	0	1	1	0	0	0	1	0

```
# examine age distributions across assignment groups

assignment_status.labs <- c("Control", "Treatment-Female Voice", "Treatment-Male Voice")
names(assignment_status.labs) <- c("C", "TF", "TM")

ggplot(final_attriters,aes(x=age))+geom_histogram(bins=30)+
  facet_grid(~assignment_status, labeller = labeller(assignment_status = assignment_status.labs))+
  labs(title = 'Histograms of age distribution by assignment group')+
  theme_bw()
```



Discussion on Non-Compliance

There are two possible forms of non-compliance: (1) not completing the full set of 5 tic-tac-toe games, and (2) choosing not to receive the treatment (i.e. not listen to the audio directive) when playing the game.

The first form of non-compliance is concerning for the experiment if subjects not completing the 5 tic-tac-toe games are not randomly related with their potential outcomes. That is, those who did not complete the games might be related to their treatment status. We do not view this a concern in the experiment as we found no subjects that did not complete all five tic-tac-toe games once initiating game play.

The second form of non-compliance here is subtle because we will not know if the participants choose to hear the directive or not (by keeping the audio down or off or listening to something else while playing), even if they complete the game. We do, however, have a screen prompt prior to game play where subjects are forced to verify that their audio is working before advancing to the game. The experiment requires experimental subjects to complete the game in isolation while online, therefore, as researchers, we have no way to tell whether the experimental subjects actually listen to the audio directives. This is a source of potential unobserved error that can undermine the observed treatment effect and a potential limitation of the study.

Discussion on Excludability

The exclusion restriction is not violated in this study since only the relevant causal agent is the receipt of the treatment. That is, potential outcomes in following suggested directives respond solely to the receipt of the treatment, not to the random assignment of the treatment. Since subjects play the online tic-tac-toe game in isolation and only receive the programmed treatment or control, subjects must only be exposed to the treatment or control. As the experiment has been designed, the same procedures are used to measure

outcomes in the treatment and control groups. No other research activities have been performed other than the treatment of interest that differentially impacts treatment and control groups.

Discussion on non-interference In our experiment, the potential outcomes $Y_i(d)$ will be assumed to be unaffected by the treatment of other subjects. Subjects are provided the treatment in isolation and play the online tic-tac-toe game without others around them. Moreover, since the suggested directives are random during game play, one subject's treatment should not affect another subject's outcomes. We can say that our experiment upholds the Stable Unit Treatment Value Assumption (SUTVA) or non-interference.

However, non-interference may be violated if people conspire and speak to each other before, or even during game play. There is a remote possibility of a non-interference violation if, for example, subjects who live in the same household or work for the same company receive treatment and discuss the game dynamics among themselves. These subjects living or working together may infer that they are receiving different treatments and change their behavior in game play from what they otherwise would have had they played in isolation. In our study, there are nine subjects belonging to the same company (United Fire Group) and three people living in the same house. Furthermore, there are also five MIDS students that may be taking W241 or may have taken W241 in the past that may infer something implicitly about the treatment. Again, these students may have the possibility of discussing the treatment among themselves. However, given that we have not communicated who the participants are in this study, we can assume that communication among treated students in the experiment is a remote possibility. We can assume no interdependencies in game play strategy and the assignment of one subject has no consequences for outcomes of other subjects.

CONSORT

The flow diagram starts with 72 subjects who signed up to participate in the study through outreach through social outlets by researchers. The 72 subjects were randomly allocated into three groups - control (no audio directive), treatment (male audio directive) and treatment (female audio directive) following a blocking design based on subject gender (male or female). Following the block randomization, roughly half of subjects were assigned to control (33 subjects) and the other half assigned to one of the treatment groups (39 subjects). Subjects assigned to treatment were randomly assigned to either the male audio directive treatment and female audio directive treatment (22 assigned to the male audio directive and 17 assigned to the female audio directive). Each subject was sent an email to play the series of tic-tac-toe games under their assigned experimental condition (control or treatment). The game play outcomes related to compliance to the directive, by move and overall compliance will be used for estimating the treatment effects. No attrition occurred in this randomized experiment.

Power Calculation

According to List et al. (2008), the power of a statistical test is the probability that it will correctly lead to the rejection of the null hypothesis (the probability of a Type II error is 1-power, and is equal to the probability of falsely not rejecting the null hypothesis). The idea behind the choice of optimal sample sizes in this scenario is that the sample sizes have to be just large enough so that the experimenter (1) does not falsely reject the null hypothesis that the population treatment and control outcomes are equal, i.e., commit a Type I error; and (2) does not falsely accept the null hypothesis when the actual difference is equal to δ, i.e. commit a Type II error. A simple rule of thumb to maximize power given a fixed experimental budget naturally follows: the ratio of the sample sizes is equal to the ratio of the standard deviations of outcomes.

In our experiment, we assume the hypothetical effect size to be achieved should be 0.5, whereby average compliance rate for those in the control is p=0.5 and average compliance rate for those in the treatment is p=1. We understand that this might be reaching a bit with this hypothetical effect size since an average compliance rate of p=1 is the most extreme upward bound. It is likely that the realized treatment effect will be smaller in the experiment. Below we discuss the effect on power using smaller effect sizes and illustrate the strength of power over effect size-sample size space.

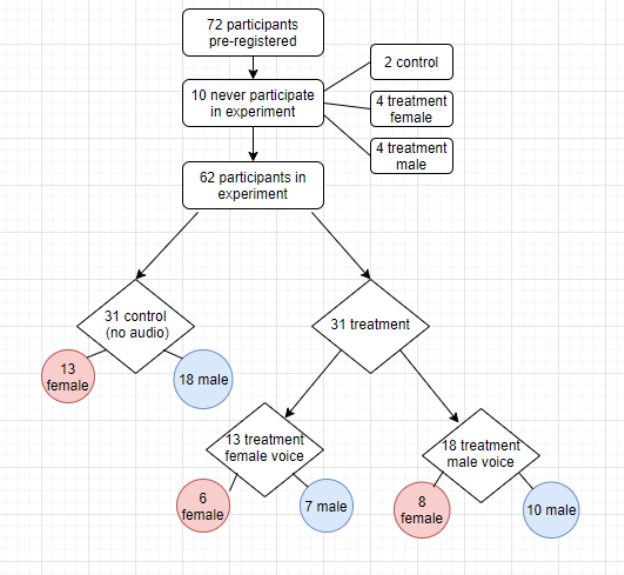


Figure 1: Random Assignment Flow Chart

We compute the appropriate sample size for the given effect size, significance level and power using the test of proportions.

```
# assume effectsize from the t_test_result
effectsize <- 0.5
# test of proportions
pwr.2p.test(h = effectsize, n = NULL , sig.level = .05, power = .6 )
```

Assuming a reasonable target power of 60%, significance level $\alpha = .05$ and effect size of 0.5, we would need sample size of 39 in both groups.

Following the pre-treatment survey, we were able to solicit participation from 72 people. Given the known sample sizes in the study, assuming a hypothetical effect size of 0.5 and significance level $\alpha = .05$, we can compute the power in the study as follows:

```
# number of observations in the control and treatment groups in the study
n_control = nrow(compliance_rates[compliance_rates$assignment_status=='C'])
n_treat = nrow(compliance_rates[compliance_rates$assignment_status!='C'])
n_treatmale = nrow(compliance_rates[compliance_rates$assignment_status=='TM'])
n_treatfemale = nrow(compliance_rates[compliance_rates$assignment_status=='TF'])

# compute the power for the overall control and combined treatment groups
pwr.2p2n.test(h = effectsize, n1 = n_control , n2 = n_treat ,
              sig.level = .05, power = NULL )

# compute the power for the overall control and male audio treatment groups
pwr.2p2n.test(h = effectsize, n1 = n_control , n2 = n_treatmale ,
              sig.level = .05, power = NULL )

# compute the power for the overall control and female audio treatment groups
pwr.2p2n.test(h = effectsize, n1 = n_control , n2 = n_treatfemale ,
              sig.level = .05, power = NULL )
```

```r
round(pwr.2p2n.test(h = effectsize, n1 = n_control , n2 = n_treat , sig.level = .05, power = NULL)$powe
```

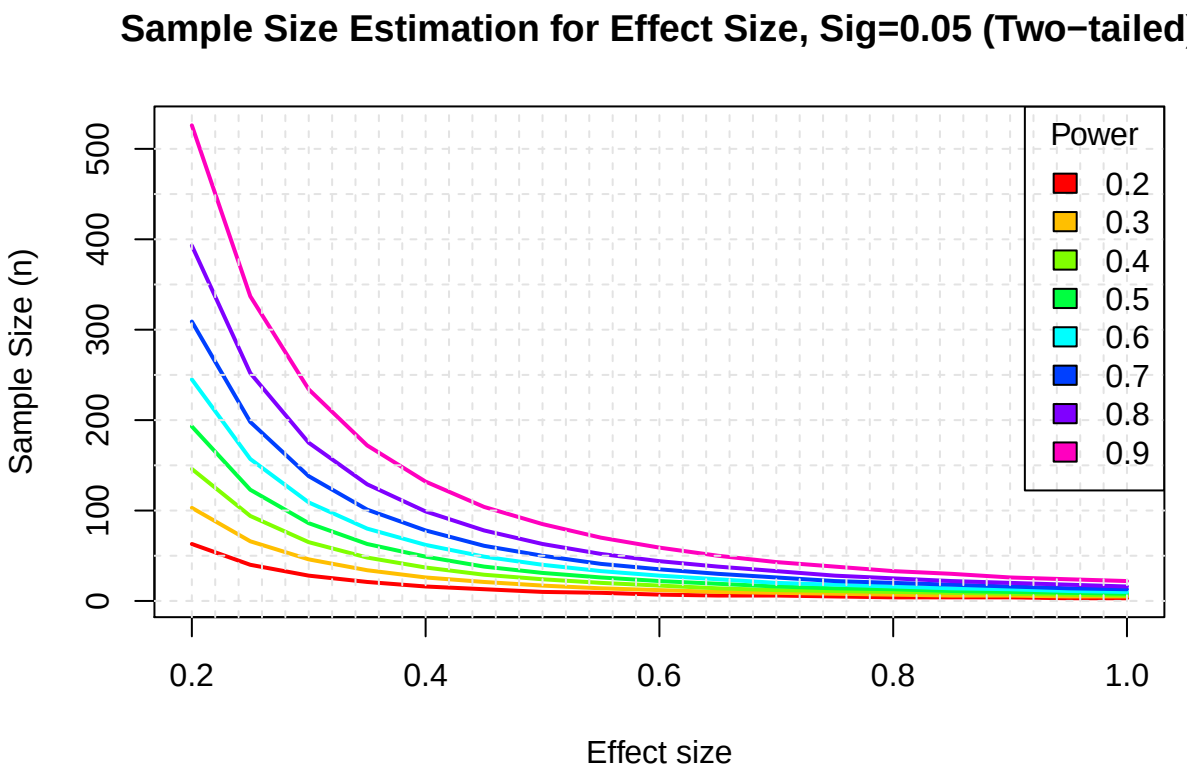
```
## [1] 60
```

Following the power of our study will be 56 percent, suggesting a moderately-powered experiment.

If we consider the studies with the individual treatment effects (male and female audio) only, and assuming our hypothetical anticipated effect size, the power would have been 44 percent and 39 percent for the male and female audio treatment studies, respectively.

As an exposition, we can visualize power curves below to understand the relationship with effect size and sample size. Intuitively, to achieve an adequately-powered small effect size, we would need larger sample sizes.

```r
# power values
p <- seq(.2,.9,.1)
np <- length(p)

# range of effect sizes
h <- seq(.2,1,.05)
nh <- length(h)

# obtain sample sizes
samsize <- array(numeric(nh*np), dim=c(nh,np))
for (i in 1:np){
  for (j in 1:nh){
    result <- pwr.2p.test(h = h[j], n = NULL , sig.level = .05, power = p[i])
    samsize[j,i] <- ceiling(result$n)
  }
}

# set up graph
xrange <- range(h)
yrange <- round(range(samsize))
colors <- rainbow(length(p))
plot(xrange, yrange, type="n",
  xlab="Effect size",
  ylab="Sample Size (n)" )

# add power curves
for (i in 1:np){
  lines(h, samsize[,i], type="l", lwd=2, col=colors[i])
}

# add annotation (grid lines, title, legend)
abline(v=0, h=seq(0,yrange[2],50), lty=2, col="grey89")
abline(h=0, v=seq(xrange[1],xrange[2],.02), lty=2,
   col="grey89")
title("Sample Size Estimation for Effect Size, Sig=0.05 (Two-tailed)")
legend("topright", title="Power", as.character(p),
   fill=colors)
```



In the actual study, we accounting for attrition, we had just 62 subjects in the experiment and therefore, could not achieve the power targeted a priori. Moreover, as we discuss in the Results section, our treatment effects were substantially smaller than 0.5. We report the actual power of the experiment in the Results sub-section "Post-treatment power calculation".

Data Collection

Subjects for our experiment were a collection of fellow students in the MIDS program, and the authors' personal and professional acquaintances. Initial data collection involved sending out a pre-treatment survey in which the respondent is asked to provide their gender, age, and contact information. We use the data collected from this survey to inform our blocking strategy, as described above in the Randomization Process section. One additional step that was included was the creation of a pilot group, which we used to identify any technical bugs in our software infrastructure. This pilot group contained 9 randomly assigned subjects with proportional representation from the full-sized treatment groups, and received treatment before the remainder of the subject pool were contacted to receive treatment.

Treatment was administered via a custom designed web application. Emails were sent to the email addresses that were provided in the pre-treatment survey that contained links to the web application that were unique to the individual subjects, which allowed us to identify which subject was connecting to the web application. If the subject had not completed their assigned task within a week prior to the deadline listed in the email, an additional reminder email is sent out. The script of both of these emails can be found in the appendix.

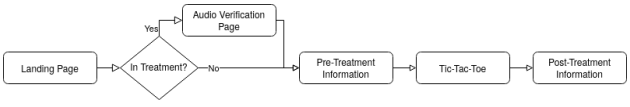


Figure 2: Web-application control flow

Figure x: Web-application control flow

Upon arrival at the website, the user is greeted with a page that informs the subject that they are expected to fully complete the game, and navigation away from the game will invalidate their results. Those in either treatment group are shown a button which allows them to test their audio. On advancing to the next page of the website, subjects in the treatment group are also greeted with an additional audio test, which requires them to select a button on screen that corresponds to a verbal direction. Subjects in the treatment group are not allowed to advance to subsequent screens if they do not pass this audio validation check. The next page then notifies the user of the game that they will play, and informs them that the move suggestions that they receive may or may not be optimal.

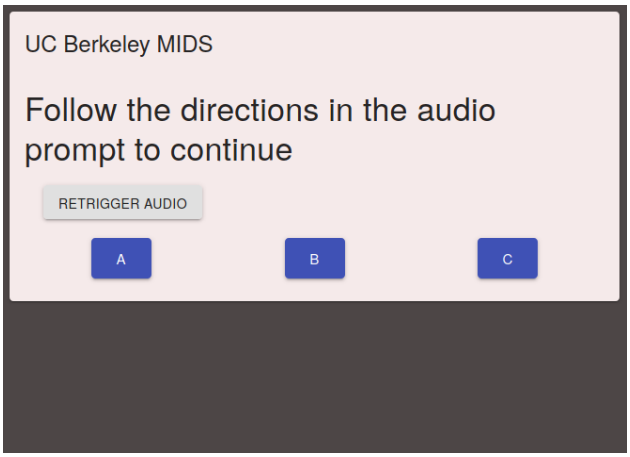


Figure 3: Screenshot of the audio verification page

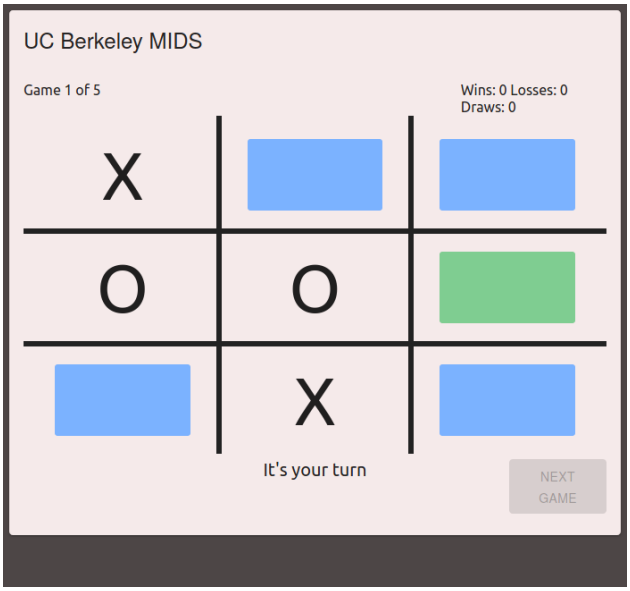


Figure 4: Screenshot of the tic-tac-toe game

Once in the game, the users were tasked with completing five games of tic-tac-toe. The opening move alternated between computer and player, with the player getting the opening move in the first game. On each of the subject's turns, a suggested cell is highlighted to a different color than the other cells. This suggested move is either (with equal probability): the optimal move as determined by a Minimax decision rule, or a randomly selected available cell on the board

This approach for suggesting a position was taken in order to discourage the subject from assuming the

suggestions are always optimal or never optimal. The computer opponent follows the same strategy when making their selection. If the subject is in either treatment group, an audio clip is played with a verbal direction to move to the highlighted cell with the gender of the speaker corresponding to the treatment group that the subject belongs to. On the player's selection of their next move the current board state, selected position, suggested position, and whether the suggested position was optimal are all recorded.

On completion of five games, the subjects are brought to a final screen which thanks them for participating and contains a link to a post-experiment survey. This survey's results were used to identify any technical glitches and inform some retrospective analysis of the experiment design.

Model

The main model of our analysis investigates the suggested move compliance rate of the subject dependent on whether the subject was in the control or one of the treatment groups. Our models can be expressed formally as:

- Linear model without interaction terms:

$Y = \beta_1 maleaudio + \beta_2 femaleaudio + \beta_3 gender + \beta_4 age$

- Linear model with interaction terms:

$Y = \beta_1 maleaudio + \beta_2 femaleaudio + \beta_3 gender + \beta_4 age + \beta_5 maleaudio * gender + \beta_6 femaleaudio * gender + \beta_7 maleaudio * age + \beta_8 femaleaudio * age$

- Logistic regression model without interaction terms:

$P(Y = 1) = \beta_1 maleaudio + \beta_2 femaleaudio + \beta_3 gender + \beta_4 age$

or expressed in log-odds as:

$LL = ln(\frac{p}{1-p}) = \beta_1 maleaudio + \beta_2 femaleaudio + \beta_3 gender + \beta_4 age$

We estimated causal effects using the postulated models over a sequence of analyses:

- 1. Proportional compliance to suggested moves for all moves
- 2. Proportional compliance to suggested moves for optimal and non-optimal suggested moves
- 3. Move-by-move compliance to suggested moves for all moves (using logistic regression)
- 4. Proportional compliance to suggested moves for all moves by subject gender
- 5. Proportional compliance to suggested moves by order of game
- 6. Proportional compliance to suggested moves by order of move

Each of these estimates causal effects in an attempt to more fully comprehend the dynamic between move suggestions and game play strategy.

Model Results

The regression table below reports the model estimates, first with no covariates, then with additional covariates, robust standard errors, and clustered standard errors We clustered on subject identifier since each subject received one treatment throughout game play.

```
# report the formatted regression results
stargazer(naive_reg, reg, reg, reg_interaction, reg_interaction,
          type = 'text',
          se=list(naive_reg$robustse, reg$robustse, reg$cluster1se,
                  reg_interaction$robustse, reg_interaction$cluster1se),
```

```
        add.lines = list(c('SE', 'Robust', 'Robust', 'Clustered',
                           'Robust', 'Clustered')),
        no.space = TRUE,align=TRUE,table.placement="H",
        header=F)
```

```
##
## =========================================================================================================
##                                                                                           Dependent variabl
##                                            ----------------------------------------------------------------
##                                                                                                 comply_rate
##                                                    (1)                    (2)                    (3)
## ---------------------------------------------------------------------------------------------------------
## as.factor(assignment_status)TF                 0.289***               0.260***               0.260***
##                                                (0.075)                (0.077)                (0.080)
## as.factor(assignment_status)TM                 0.203***               0.202***               0.202***
##                                                (0.061)                (0.057)                (0.060)
## genderM                                                               0.059                  0.059
##                                                                       (0.050)                (0.053)
## age                                                                   -0.004**               -0.004*
##                                                                       (0.002)                (0.002)
## as.factor(assignment_status)TF:genderM
##
## as.factor(assignment_status)TM:genderM
##
## as.factor(assignment_status)TF:age
##
## as.factor(assignment_status)TM:age
##
## Constant                                       0.403***               0.525***               0.525***
##                                                (0.030)                (0.092)                (0.096)
## ---------------------------------------------------------------------------------------------------------
## SE                                              Robust                 Robust                Clustered
## Observations                                      62                     62                     62
## R2                                              0.268                  0.329                  0.329
## Adjusted R2                                     0.243                  0.282                  0.282
## Residual Std. Error                        0.209 (df = 59)        0.204 (df = 57)        0.204 (df = 5
## F Statistic                           10.781*** (df = 2; 59) 6.979*** (df = 4; 57) 6.979*** (df = 4
## =========================================================================================================
## Note:
```

From the above plot, we see that voice directives have a statistically significant and positive effect on the willingness of a subject to take the move that was directed. We also see that the female speaker's voice has a stronger effect on subject move compliance. The female voice treatment has coefficient estimate of 0.2595786 (with robust standard error of 0.0804427). The male voice treatment has a coefficient estimate of 0.2023633 (with robust standard error of 0.0597742).The coefficient for the age variable does indicate a very small, yet statistically significant, negative effect on the willingness to listen. The coefficient for gender is positive, but insignificant.

However, after including the interaction terms between gender and the treatment assignments and age and the treatment assignments, the coefficients for the treatments are no longer statistically significant. We observe that the standard errors increase with the inclusion of the interaction terms due to the collinearity between the interaction terms and the stand-alone covariates, male and female treatment assignment, gender and age. This is because the residual variance falls only slightly with the inclusion of the interaction terms, while the standard deviation of $\tilde{X}_{ki}$ increases (the residual of the regression of X_{ki} on all other regressors). Only the coefficient on the `age` variable remains statistically significant across all regression specifications,

although it is practically insignificant in size.

We conducted a two-sample proportional t-test to observe that the difference between male and female voice directives is not statistically significant. We caution that the sample sizes are too small to render any meaningfully power to these effects. See the R code below for the t-test.

```
# conduct a two-sample t-test and save results to t_test_result variable
t_test_result_tm <- t.test(final_tm, final_c)
t_test_result_tm
```

```
##
##  Welch Two Sample t-test
##
## data:  final_tm and final_c
## t = 3.2568, df = 27.815, p-value = 0.002963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07536913 0.33110126
## sample estimates:
## mean of x mean of y
## 0.6058984 0.4026632
```

```
t_test_result_tf <- t.test(final_tf, final_c)
t_test_result_tf
```

```
##
##  Welch Two Sample t-test
##
## data:  final_tf and final_c
## t = 3.7016, df = 16.506, p-value = 0.001849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1238951 0.4540811
## sample estimates:
## mean of x mean of y
## 0.6916513 0.4026632
```

```
t_test_result_t <- t.test(final_tm, final_tf)
t_test_result_t
```

```
##
##  Welch Two Sample t-test
##
## data:  final_tm and final_tf
## t = -0.95128, df = 24.116, p-value = 0.3509
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2717551  0.1002493
## sample estimates:
## mean of x mean of y
## 0.6058984 0.6916513
```

Consistent with the regression results, we observe that the treatment effect for the male voice and female voice treatments are statistically different from zero with a p-value in the t-test of 0.003 and 0.0018, respectively. The effect between the gendered voice treatments is statistically insignificant with a p-value in the t-test of 0.3509.

Evaluation of compliance with suggestions on a move-by-move basis

We examined the effects of a speaker giving voice directives on a move-by-move basis. Instead of the outcome being the proportion of all moves complied with, the outcome in this analysis is the binary outcome of compliance (0 or 1) for a given move suggestion. We estimated a logistic regression of the binary comply outcome on the treatment assignment status adjusting for the subject gender and age covariates and with subject id-level fixed effects might tell a more nuanced story. Report results with both robust standard errors and clustering at the subject ID level. Clustering at the subject ID level is important since the number of the moves (and therefore, number of observations) may differ across subjects. Clustering ensures that the standard errors reflects the tighter variation introduced at the subject identifier level. The table below shows the logistic regression results reporting coefficients as logits.

```
# report the formatted regression results
stargazer(logistic_reg(data=all_moves), logistic_reg(data=all_moves),
          type = 'text',
          se = list(robust_se(logistic_reg(data=all_moves)),
                    cluster_se(logistic_reg(data=all_moves),all_moves)),
          add.lines = list(c('Fixed Effects','Yes','Yes'),
                           c('SE', 'Robust', 'Clustered')),
          omit.stat = c('ser', 'F'),
          omit = c("subject_id"),
          no.space = TRUE,align=TRUE,table.placement="H",
          header=F)
```

```
##
## ================================================================================
##                                          Dependent variable:
##                                  ---------------------------------------------
##                                                    comply
##                                          (1)                    (2)
## --------------------------------------------------------------------------------
## as.factor(assignment_status)TF 22,636,711,162,643.000 22,636,711,162,643.000
##
## as.factor(assignment_status)TM         0.416                  0.416
##                                                             (2,067.895)
## genderM                                -0.032                 -0.032
##
## age                                    -0.073                 -0.073
##
## Constant                               2.515                  2.515
##
## --------------------------------------------------------------------------------
## Fixed Effects                           Yes                    Yes
## SE                                     Robust                Clustered
## Observations                           1,166                  1,166
## Log Likelihood                        -655.727               -655.727
## Akaike Inf. Crit.                    1,437.453              1,437.453
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

Exponentiating the coefficients, we can interpret the likelihoods of compliance with suggested moves as odds-ratios. We find that the female audio treatment effect increases odds of compliance with a suggested move by an infinite factor ∞, controlling for covariates and fixed effects. The male audio treatment effect increases odds of compliance with a suggested move by a factor of 1.52. Being a male subject increases odds of compliance with a suggested move by a factor of 0.97.

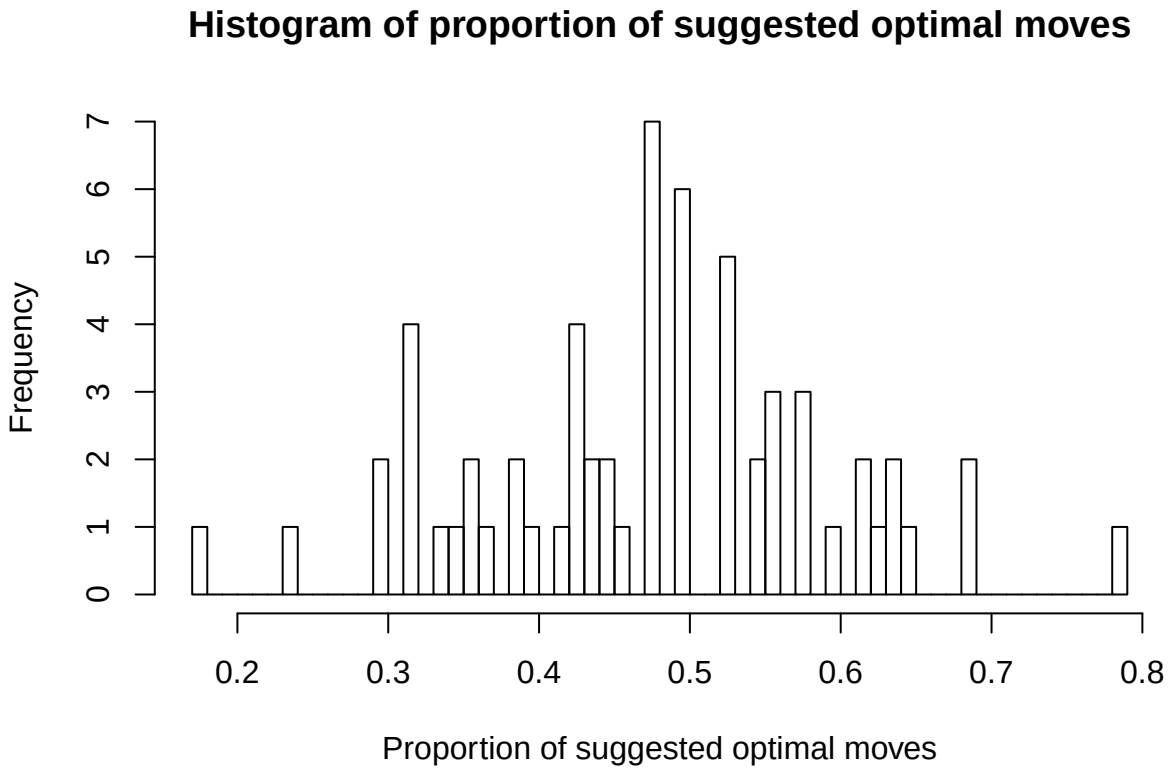
This model appears to reinforce the findings of the previous model, which is that audio directives are much more likely to be followed than visual directives. Female speakers has a stronger positive effect (larger odds of eliciting compliance with the suggested move directive).

Evaluation of compliance to suggested optimal verses non-optimal moves

As discussed in the Experimental Design section, subjects randomly received optimal and non-optimal move suggestions throughout game play. In order to assess whether the main results were consistent for the optimal and non-optimal move suggestions, we estimated the treatment effects on the subset of the data corresponding to the two type of move suggestions.

First, as a diagnostic check, we look at the distribution of the proportion of optimal moves received by each subject. Some subject received more optimal moves than others, but distribution of proportion of optimal suggested moves by subject should be broadly symmetric since move suggestions were random.

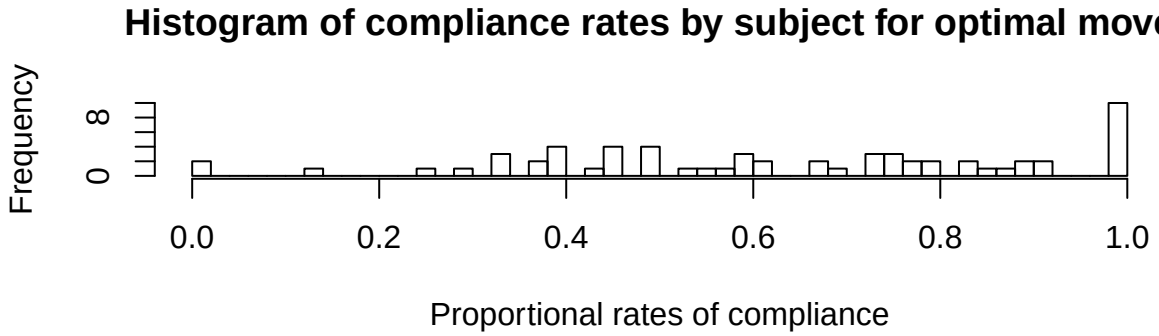
```
hist(all_moves_agg$mean_optimal_move, main="Histogram of proportion of suggested optimal moves",
     xlab="Proportion of suggested optimal moves", ylab = "Frequency", breaks=50)
```

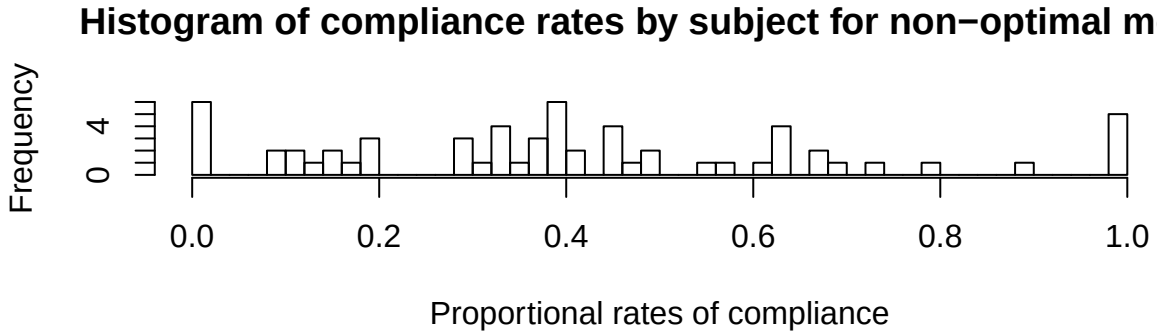


Second, as further diagnostic check, the histograms below show the distribution of the proportion of compliance to optimal verses non-optimal suggested moves. The distribution is more right-skewed for the optimal moves than for non-optimal moves. As we would expect a priori, there is a tendency for higher rates of compliance to optimal moves than non-optimal suggested moves.

```
par(mfrow=c(2,1))

# show distribution of compliance rates by subject
hist(optimal_moves_agg$comply_rate,
     main="Histogram of compliance rates by subject for optimal moves",
     xlab="Proportional rates of compliance", ylab = "Frequency", breaks=50)
```

```r
# show distribution of non-compliance rates by subject
hist(nonoptimal_moves_agg$comply_rate,
     main="Histogram of compliance rates by subject for non-optimal moves",
     xlab="Proportional rates of compliance", ylab = "Frequency", breaks=50)
```





We find the compliance to suggested optimal moves are higher across all treatment groups. See the tables for the compliance rates by assignment status below.

```r
par(mfrow=c(1,2))
# report compliance rates across all treatment groups
print("Proportional compliance rates for optimal moves")
```

```
## [1] "Proportional compliance rates for optimal moves"
```

```r
kable(complyratesbytreatment(dt=all_moves_optimal))
```

assignment_status	mean_comply_rate
C	0.5404412
TF	0.7931034
TM	0.7065868

```r
print("Proportional compliance rates for non-optimal moves")
```

```
## [1] "Proportional compliance rates for non-optimal moves"
```

```r
kable(complyratesbytreatment(dt=all_moves_nonoptimal))
```

assignment_status	mean_comply_rate
C	0.2977346
TF	0.6015038
TM	0.5207101

However, the estimated treatment effects are larger for the non-optimal move suggestions. This gives evidence

that people are differentially listening to the audio when making their moves strategically over the course of play.

```
#complyratesbytreatment(dt=all_moves_optimal)
par(mfrow=c(3,2))
print('ATE for Male Treatment over optimal moves:')
```

```
## [1] "ATE for Male Treatment over optimal moves:"
```

```
round(ate_treatmentmale(dt=all_moves_optimal),3)
```

```
## [1] 0.166
```

```
print('ATE for Male Treatment over nonoptimal moves:')
```

```
## [1] "ATE for Male Treatment over nonoptimal moves:"
```

```
round(ate_treatmentmale(dt=all_moves_nonoptimal),3)
```

```
## [1] 0.223
```

```
print('ATE for Female Treatment over optimal moves:')
```

```
## [1] "ATE for Female Treatment over optimal moves:"
```

```
round(ate_treatmentfemale(dt=all_moves_optimal),3)
```

```
## [1] 0.253
```

```
print('ATE for Female Treatment over nonoptimal moves:')
```

```
## [1] "ATE for Female Treatment over nonoptimal moves:"
```

```
round(ate_treatmentfemale(dt=all_moves_nonoptimal),3)
```

```
## [1] 0.304
```

```
kable(ate_treatmentmale(dt=all_moves_optimal))
```

x
0.1661456

```
kable(ate_treatmentmale(dt=all_moves_nonoptimal))
```

x
0.2229754

```
kable(ate_treatmentfemale(dt=all_moves_optimal))
```

x
0.2526623

```
kable(ate_treatmentfemale(dt=all_moves_nonoptimal))
```

x
0.3037691

```
kable(ate_treatment_btwnmalefemale(dt=all_moves_optimal))
```

x
-0.0865166

```
kable(ate_treatment_btwnmalefemale(dt=all_moves_nonoptimal))
```

x
-0.0807937

Below we report the linear model estimates in two separate tables for optimal and non-optimal moves. Each table shows estimates in the first column with gender and age covariates reporting robust standard errors, second column with clustered standard errors (clustering on subject identifier), third column with interaction terms reporting robust standard errors, and fourth column with interaction terms with clustered standard errors (clustering on subject identifier).

```
# report the formatted regression results
stargazer(reg_optimal, reg_optimal, reg_interaction_optimal, reg_interaction_optimal,
          type = 'text',
          se = list(robust_se(reg(data=optimal_moves_agg)),
                    cluster_se(reg(data=optimal_moves_agg),optimal_moves_agg),
                    robust_se(reg_interaction(data=optimal_moves_agg)),
                    cluster_se(reg_interaction(data=optimal_moves_agg),optimal_moves_agg)),
          add.lines = list(c('Move Type', 'Optimal', 'Optimal', 'Optimal', 'Optimal'),
                           c('SE', 'Robust', 'Clustered', 'Robust', 'Clustered')),
          omit.stat = c('ser', 'F'),
          no.space = TRUE,align=TRUE,table.placement="H",
          header=F)
```

```
##
## ================================================================================
##                                                Dependent variable:
##                                            -----------------------------------
##                                                       comply_rate
##                                              (1)       (2)      (3)       (4)
## --------------------------------------------------------------------------------
## as.factor(assignment_status)TF             0.195**   0.195**   0.288     0.288
##                                            (0.085)   (0.081)  (0.465)   (0.350)
## as.factor(assignment_status)TM             0.152**   0.152**   0.045     0.045
##                                            (0.072)   (0.068)  (0.435)   (0.314)
## genderM                                     -0.011   -0.011    -0.122   -0.122
##                                            (0.064)   (0.060)  (0.095)   (0.092)
## age                                        -0.007** -0.007***  -0.006  -0.006**
##                                            (0.003)   (0.003)  (0.004)   (0.003)
## as.factor(assignment_status)TF:genderM                         0.232     0.232
##                                                               (0.178)   (0.155)
## as.factor(assignment_status)TM:genderM                         0.224    0.224*
##                                                               (0.147)   (0.135)
## as.factor(assignment_status)TF:age                             -0.007   -0.007
##                                                               (0.015)   (0.011)
## as.factor(assignment_status)TM:age                             -0.001   -0.001
##                                                               (0.012)   (0.008)
## Constant                                   0.830*** 0.830***  0.854*** 0.854***
##                                            (0.138)   (0.122)  (0.161)   (0.126)
## --------------------------------------------------------------------------------
## Move Type                                  Optimal   Optimal  Optimal   Optimal
## SE                                          Robust  Clustered  Robust  Clustered
## Observations                                  62        62       62        62
## R2                                          0.264     0.264    0.314     0.314
## Adjusted R2                                 0.212     0.212    0.210     0.210
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

```
# report the formatted regression results
stargazer(reg(data=nonoptimal_moves_agg), reg(data=nonoptimal_moves_agg),
```

```
          reg_interaction(data=nonoptimal_moves_agg), reg_interaction(data=nonoptimal_moves_agg),
          type = 'text',
          se = list(robust_se(reg(data=nonoptimal_moves_agg)),
                    cluster_se(reg(data=nonoptimal_moves_agg),nonoptimal_moves_agg),
                    robust_se(reg_interaction(data=nonoptimal_moves_agg)),
                    cluster_se(reg_interaction(data=nonoptimal_moves_agg),nonoptimal_moves_agg)),
          add.lines = list(c('Move Type', 'Nonoptimal', 'Nonoptimal', 'Nonoptimal', 'Nonoptimal'),
                           c('SE', 'Robust', 'Clustered', 'Robust', 'Clustered')),
          omit.stat = c('ser', 'F'),
          no.space = TRUE,align=TRUE,table.placement="H",
          header=F)
```

```
##
## ================================================================================
##                                                Dependent variable:
##                                    ------------------------------------------
##                                                     comply_rate
##                                       (1)        (2)        (3)        (4)
## --------------------------------------------------------------------------------
## as.factor(assignment_status)TF      0.333***   0.333***    0.160      0.160
##                                     (0.101)    (0.096)    (0.669)    (0.485)
## as.factor(assignment_status)TM      0.226***   0.226***    0.052      0.052
##                                     (0.070)    (0.067)    (0.467)    (0.356)
## genderM                              0.117*     0.117*     0.057      0.057
##                                     (0.064)    (0.061)    (0.075)    (0.073)
## age                                  -0.002     -0.002     -0.002     -0.002
##                                     (0.003)    (0.002)    (0.003)    (0.003)
## as.factor(assignment_status)TF:genderM                      0.127      0.127
##                                                            (0.230)    (0.188)
## as.factor(assignment_status)TM:genderM                      0.139      0.139
##                                                            (0.206)    (0.176)
## as.factor(assignment_status)TF:age                          0.003      0.003
##                                                            (0.025)    (0.018)
## as.factor(assignment_status)TM:age                          0.003      0.003
##                                                            (0.010)    (0.007)
## Constant                            0.273**    0.273**    0.323**    0.323***
##                                     (0.118)    (0.107)    (0.150)    (0.118)
## --------------------------------------------------------------------------------
## Move Type                         Nonoptimal Nonoptimal Nonoptimal Nonoptimal
## SE                                  Robust   Clustered    Robust   Clustered
## Observations                          62         62         62         62
## R2                                   0.325      0.325      0.339      0.339
## Adjusted R2                          0.277      0.277      0.239      0.239
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

When breaking out comparison down by the optimal and non-optimal moves, we still see that the directives delivered by audio lead to a higher and statistically significant move compliance rate regardless of whether the move is optimal or not. This further supports our initial findings that compliance rates differ between audio and visual directions. We note that the effect is stronger in size for the non-optimal moves for both gendered voices, suggesting that subjects listen regardless of game strategy Compare the size of the coefficient estimates for the female and male audio treatments, respectively, (0.1954951 and 0.1523081) for optimal moves verses nonoptimal moves (0.332945 and 0.2264379).

When examining the comparison between the speaker's gender, we see again that there is a slightly higher

treatment effect for the female speaker, but the difference in effect sizes is not statistically significant following a two-sample t-test.

```
t_test_result_optimal_tm
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = 3.5704, df = 375.24, p-value = 0.0004027
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07464582 0.25764548
## sample estimates:
## mean of x mean of y
## 0.7065868 0.5404412
```

```
t_test_result_optimal_tf
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = 5.2194, df = 263.96, p-value = 3.638e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1573471 0.3479775
## sample estimates:
## mean of x mean of y
## 0.7931034 0.5404412
```

```
t_test_result_optimal_t
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = -1.6725, df = 264.2, p-value = 0.0956
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.18836847  0.01533523
## sample estimates:
## mean of x mean of y
## 0.7065868 0.7931034
```

```
t_test_result_nonoptimal_tm
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = 4.7928, df = 320.16, p-value = 2.522e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1314458 0.3145051
## sample estimates:
## mean of x mean of y
```

```
## 0.5207101 0.2977346
```

```
t_test_result_nonoptimal_tf
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = 6.0818, df = 235.06, p-value = 4.775e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2053673 0.4021709
## sample estimates:
## mean of x mean of y
## 0.6015038 0.2977346
```

```
t_test_result_nonoptimal_t
```

```
##
##  Welch Two Sample t-test
##
## data:  optimal_t and optimal_c
## t = -1.4061, df = 285.95, p-value = 0.1608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.19388809  0.03230069
## sample estimates:
## mean of x mean of y
## 0.5207101 0.6015038
```

Consistent with the regression results, from the t-tests we observe that the treatment effect for the male voice and female voice treatments for the optimal moves are statistically different from zero with a p-value in the t-test of 4×10^{-4} and 0, respectively. We observe similar effects from the t-tests for the non-optimal moves with p-values of 0 and 0, respectively. The effect between the gendered voice treatments for both optimal and non-optimal moves is statistically insignificant with a p-value in the t-test of 0.0956 and 0.1608.

Evaluation of compliance to suggested moves based on gender of subject

We block randomized subjects into treatment assignments by their gender. This design has the benefit of allowing us to estimate treatment effects for each block. Therefore, we estimated the treatment effects on the subset of the data corresponding to the male and female subjects separately. Below, we report the regression table from our postulated model where the outcome remains the proportional compliance rates.

$Y = \beta_1 maleaudio + \beta_2 femaleaudio + \beta_3 age$

```
stargazer(reg_male, reg_male,reg_female,reg_female,
          type = 'text',
          se = list(robust_se(reg_male),
                    cluster_se(reg_male,all_moves_male_agg),
                    robust_se(reg_female),
                    cluster_se(reg_female,all_moves_female_agg)),
          add.lines = list(c('Subject Gender','Male','Male','Female','Female'),
                           c('SE', 'Robust', 'Clustered','Robust','Clustered')),
          omit.stat = c('ser', 'F'),
          header=F)
```

```
##
```

```
## ========================================================================
##                                        Dependent variable:
##                                    -----------------------------------
##                                                comply_rate
##                                      (1)       (2)      (3)       (4)
## ------------------------------------------------------------------------
## as.factor(assignment_status)TF 0.358*** 0.358***   0.138     0.138
##                                (0.118)   (0.108)  (0.130)   (0.116)
##
## as.factor(assignment_status)TM 0.294*** 0.294***   0.114     0.114
##                                (0.080)   (0.077)  (0.103)   (0.093)
##
## age                             -0.001   -0.001    -0.006   -0.006
##                                (0.002)   (0.002)  (0.005)   (0.004)
##
## Constant                       0.444*** 0.444***  0.671*** 0.671***
##                                (0.096)   (0.079)  (0.189)   (0.157)
##
## ------------------------------------------------------------------------
## Subject Gender                   Male     Male     Female   Female
## SE                              Robust  Clustered  Robust  Clustered
## Observations                      35       35        27       27
## R2                              0.447     0.447    0.225     0.225
## Adjusted R2                     0.394     0.394    0.124     0.124
## ========================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

We find that the treatment effects of the gendered audio voices are statistically significant for the male subjects only. For the male subjects, the treatment effect from the female audio voice is 0.358 with robust standard error of 0.09, while for the male audio voice the estimated treatment effect is 0.294 with robust standard error of 0.079.

For the female subjects, the treatment effects are statistically insignificant. The effect from female audio voice is 0.138 with robust standard error of 0.107, while for the male audio voice the estimated treatment effect is 0.1142745 with robust standard error of 0.095.

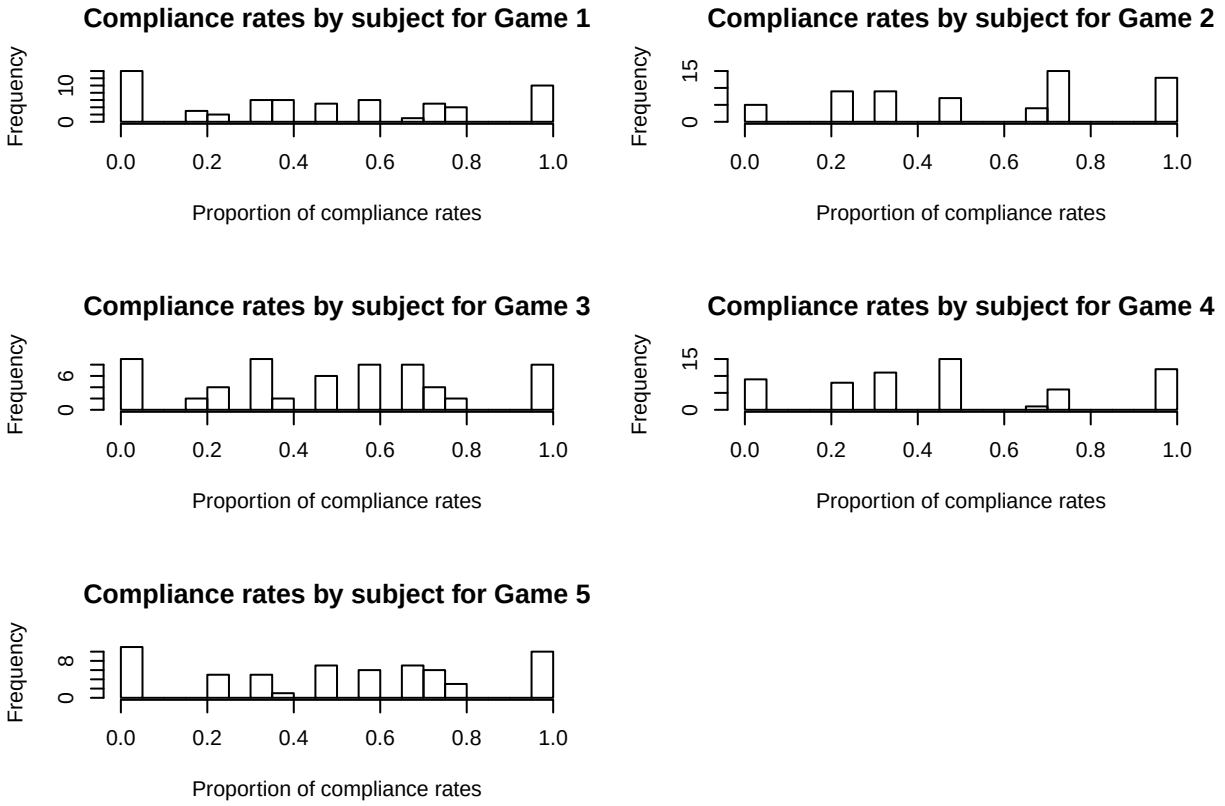
Evaluation of compliance to suggested moves based on order of game

As an ancillary analysis to assess how the game strategy interplays with the audio directives, we estimated the treatment effects by game number. This would theoretically allow us to observe whether there is a trend in compliance rates as the number of games played increases. For instance, compliance may be higher at game 1 than game 3, 4 or 5.

As a diagnostic check, we observe that the distribution of proportional compliance to suggested moves are roughly balanced by order of game.

```
# show distribution of compliance rates for each game by subject
par(mfrow=c(3,2))

hist(all_moves_game0_agg$comply_rate,
     main="Compliance rates by subject for Game 1",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=20)

hist(all_moves_game1_agg$comply_rate,
     main="Compliance rates by subject for Game 2",
```

```
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=20)

hist(all_moves_game2_agg$comply_rate,
     main="Compliance rates by subject for Game 3",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=20)

hist(all_moves_game3_agg$comply_rate,
     main="Compliance rates by subject for Game 4",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=20)

hist(all_moves_game4_agg$comply_rate,
     main="Compliance rates by subject for Game 5",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=20)
```



Estimating the linear regressions on aggregated data for all 5 games with clustered standard errors (clustering by subject identifier), we see that the treatment effects for the gendered audio voices are persistent across game order. We find statistically significant estimates for both gendered audio voice treatments over each game number. The exception is the treatment male audio voice effect in Game 5, which we ascribe as due to random chance. Over each game, the effect size is always larger for the female audio voice treatment.

```
# report the formatted regression results for all games with clustered standard errors.
stargazer(reg(all_moves_game0_agg), reg(all_moves_game1_agg),
          reg(all_moves_game2_agg),reg(all_moves_game3_agg),
          reg(all_moves_game4_agg),
          type = 'text',
          se = list(cluster_se(reg(data=all_moves_game0_agg),all_moves_game0_agg),
                    cluster_se(reg(data=all_moves_game1_agg),all_moves_game1_agg),
                    cluster_se(reg(data=all_moves_game2_agg),all_moves_game2_agg),
```

```
                    cluster_se(reg(data=all_moves_game3_agg),all_moves_game3_agg),
                    cluster_se(reg(data=all_moves_game4_agg),all_moves_game4_agg)),
          add.lines = list(c('Games','Game1','Game2','Game3','Game4','Game5'),
                           c('SE', 'Clustered', 'Clustered','Clustered','Clustered','Clustered')),
          omit.stat = c('ser', 'F'),
          header=F)
```

```
##
## ================================================================================
##                                          Dependent variable:
##                                    ---------------------------------------------
##                                                  comply_rate
##                                       (1)       (2)       (3)       (4)       (5)
## --------------------------------------------------------------------------------
## as.factor(assignment_status)TF  0.292**  0.293***   0.216*   0.286***   0.220**
##                                 (0.126)   (0.100)   (0.114)   (0.099)   (0.107)
##
## as.factor(assignment_status)TM 0.254***   0.208**   0.160**   0.224**    0.146
##                                 (0.094)   (0.089)   (0.079)   (0.096)   (0.089)
##
## genderM                          0.011     0.060    -0.020     0.104    0.141*
##                                 (0.082)   (0.076)   (0.076)   (0.079)   (0.076)
##
## age                             -0.003    -0.003    -0.004    -0.003   -0.007***
##                                 (0.003)   (0.004)   (0.003)   (0.003)   (0.003)
##
## Constant                        0.446***  0.529***  0.580***  0.424***  0.627***
##                                 (0.144)   (0.170)   (0.144)   (0.146)   (0.136)
##
## --------------------------------------------------------------------------------
## Games                            Game1     Game2     Game3     Game4     Game5
## SE                             Clustered Clustered Clustered Clustered Clustered
## Observations                      62        62        62        62        61
## R2                               0.183     0.203     0.146     0.213     0.242
## Adjusted R2                      0.126     0.147     0.086     0.157     0.188
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

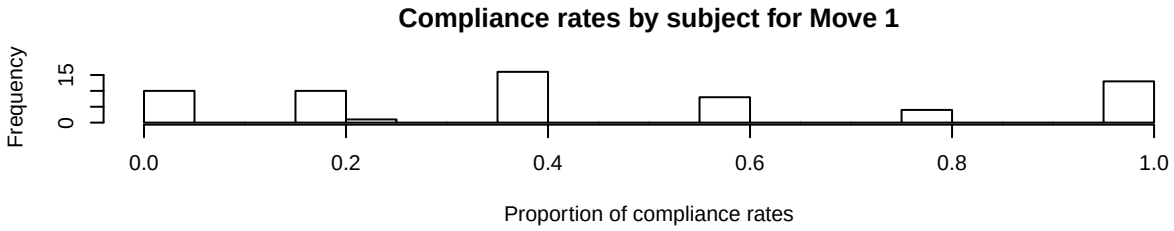
Evaluation of compliance to suggested moves based on order of move

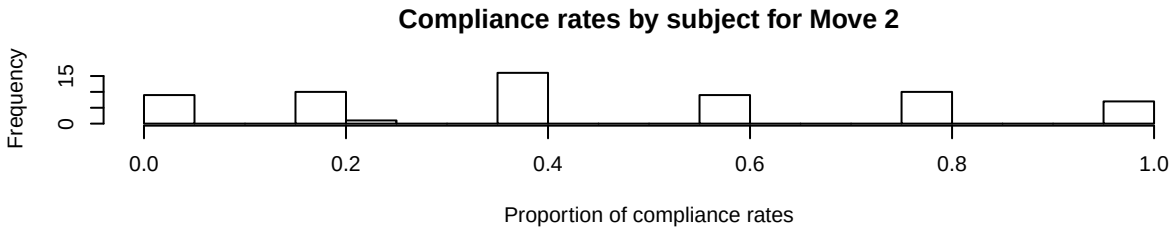
As a further ancillary analysis to assess how the game strategy interplays with the audio directives, we estimated the treatment effects by move number. This would theoretically allow us to observe whether there is a trend in compliance rates as the number of moves played increases within any given game. For instance, compliance may be higher at move 1 when there is less strategic importance to complying or not to a random suggestion than compliance in a second or third move. We assessed the causal treatment effects over the first three moves only, as all subjects would have played at least three moves in any given game.

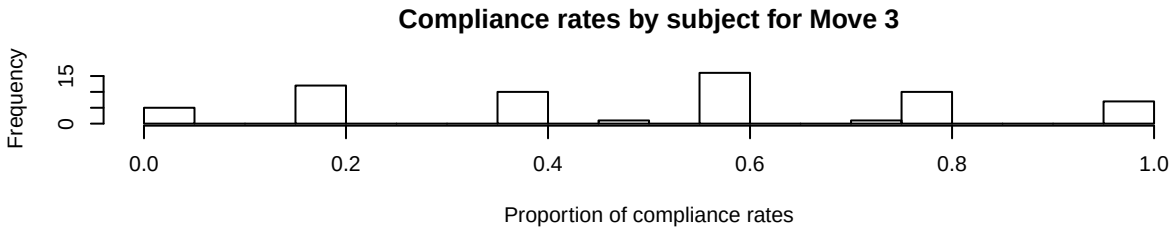
As a diagnostic check, we observe that the distribution of proportional compliance to suggested moves are roughly balanced by order of move.

```
# show distribution of compliance rates for each game by subject
par(mfrow=c(3,1))

hist(all_moves_move1_agg$comply_rate,
```

```
     main="Compliance rates by subject for Move 1",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=30)

hist(all_moves_move2_agg$comply_rate,
     main="Compliance rates by subject for Move 2",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=30)

hist(all_moves_move3_agg$comply_rate,
     main="Compliance rates by subject for Move 3",
     xlab="Proportion of compliance rates", ylab = "Frequency", breaks=30)
```







Estimating the linear regressions on aggregated data for the first 3 ordered moves with clustered standard errors (clustering by subject identifier), we see that the treatment effects for the gendered audio voices are persistent across move order. We find statistically significant estimates for both gendered audio voice treatments over each of the first three ordered moves. Over each move, the effect size is always larger for the female audio voice treatment. Curiously, the treatment effect size dwindle smaller for the female audio voice treatment over the 3 ordered moves, yet no pattern is found for the male audio voice treatments over the 3 ordered moves.

```
# report the formatted regression results
stargazer(reg(all_moves_move1_agg), reg(all_moves_move2_agg),reg(all_moves_move3_agg),
          type = 'text',
          se = list(cluster_se(reg(all_moves_move1_agg),all_moves_move1_agg),
                    cluster_se(reg(all_moves_move2_agg),all_moves_move2_agg),
                    cluster_se(reg(all_moves_move3_agg),all_moves_move3_agg)),
          add.lines = list(c('Move Order No.','Move 1','Move 2','Move 3'),
                           c('SE', 'Clustered', 'Clustered','Clustered')),
          omit.stat = c('ser', 'F'),
          header=F)
```

```
##
## ===============================================================
##                                     Dependent variable:
##                                 ----------------------------
##                                           comply_rate
##                                    (1)       (2)       (3)
## ---------------------------------------------------------------
## as.factor(assignment_status)TF 0.365***  0.299***   0.268**
##                                 (0.107)   (0.094)   (0.105)
##
## as.factor(assignment_status)TM  0.220**  0.287***   0.155**
##                                 (0.103)   (0.080)   (0.072)
##
## genderM                          0.066     0.105    -0.021
##                                 (0.081)   (0.067)   (0.069)
##
## age                             -0.001   -0.007*** -0.006**
##                                 (0.004)   (0.003)   (0.003)
##
## Constant                        0.340**  0.525***  0.647***
##                                 (0.164)   (0.128)   (0.129)
##
## ---------------------------------------------------------------
## Move Order No.                   Move 1    Move 2    Move 3
## SE                              Clustered Clustered Clustered
## Observations                       62        62        62
## R2                               0.205     0.369     0.234
## Adjusted R2                      0.149     0.324     0.181
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Result

The study does not have sufficient power to answer our initial research question of whether the gender of a speaker will affect the willingness of a listener to follow their direction. However, we do see a strong sustained effect of voice directives over visual directives on a subject's willingness to comply with suggestions. This is supported by our observation of consistently higher compliance rates among those in the treatment compared to those in the control.

We find that the effect of the female speaker is higher than the effect of the male speaker, albeit not a statistically significant difference. It is clear from our experiment that a follow up experiment should be conducted in order to examine this hypothesis more closely. The follow up experiment should include a much larger subject pool and should also include more rigid data collection strategies such that we reduce the number of attriters in the study. *Story is: We don't have sufficient power to differentiate between the effects of gender on compliance, but we do see that there is a strong sustained effect of voice directives*

Voice directions have a significant, positive effect (both male and female voice directives) in game play. We find a significant treatment effect in overall comply rates with suggested moves taken and persistence in observing these treatment effects throughout continual game play, on a by-move and by-game basis. While the effect is slightly stronger for the female voice treatment, we cannot conclude that there is a gender bias in voice directives. The difference in comply rates with suggested moves between male and female voice

treatment is not distinguishable from zero. Moreover, our study is too low-powered to provide credible evidence of the existence of gendered voice bias one way or the other.

Post-treatment power calculation

We find that the power of the observed treatment effects are considerably lower than anticipated prior to the experiment. This is due both to the smaller observed effect size relative to the theoretically anticipated effect size and the small sample sizes.

The power of our effects are 10 percent and 12 percent for the male and female audio treatment studies, respectively.

The sample sizes required to achieve 50% power for the observed effect size of the male audio voice treatment was 188 and observed effect size of the female audio voice treatment is 114. This represents a roughly 10 and 9 fold increase in the subject pool we actually had in our experiment.

Conclusions

Appendix

##Pre-treatment survey What is your gender? [Multiple Choice] - Male - Female What is your age? [Free-entry] What is a good contact email? [Free-entry] Do you consent to taking part in a study that will require you to play a game through a browser based web application with the use of audio (Approximately 10 minutes)? [Multiple Choice] - Yes - No

Post-treatment survey contents

Did you experience any technical issues? [Multiple Choice] - Yes - No If you experienced technical issues please describe below: [Free entry] Was the audio clear? [Multiple Choice] How long did it take you (in minutes) to complete the 5 games? If you exited the game early please write NA. [Free-entry] What is a good contact email? [Free-entry]

Email scripts

Initial contact

Hello, and thank you again for agreeing to take part in our study.

Below is a link to the web application where you will play a series of games, which should take no longer than 10 minutes. Please click the link when you are able to complete the game on a computer using Chrome or Firefox in a single session, as you will not be given another attempt to participate if you leave the webpage early. You will also not be allowed to proceed if you attempt to access the page on a mobile device. You must complete this task by {due_date}.

{external_link}

Thank you,

UC Berkeley students School of Information

Reminder

Hello,

Our records indicate that you have not yet completed the tasks that have been sent to you. Please navigate to the link below and complete the assigned game by {due_date}.

{external_link}

Thank you,

UC Berkeley students School of Information Web application scripts

Landing page

Treatment

Hello, and thank you for agreeing to take part in our study. You will be asked to play 5 short and simple games. It is a requirement that you have headphones or speakers, as there are some audio cues during the course of this activity. You can use the button below to test out the sound on your device. Once you are ready to start playing, please click the START button below. NOTE: If you reload or navigate away from this page before completing all games, your results will be invalidated and you will not be given another opportunity to complete all the games. Please refrain from leaving this webpage without first completing the games.

Control

Hello, and thank you for agreeing to take part in our study. You will be asked to play 5 short and simple games. Once you are ready to start playing, please click the START button below. NOTE: If you reload or navigate away from this page before completing all games, your results will be invalidated and you will not be given another opportunity to complete all the games. Please refrain from leaving this webpage without first completing the games.

Pre-treatment information page

Treatment

You will be asked to play 5 simple games of tic-tac-toe. Before each of your moves, you will hear a speaker suggest a move. These moves may or may not be optimal and you are not required to follow their suggestion. Have fun playing!

Control

You will be asked to play 5 simple games of tic-tac-toe. Before each of your moves, you will see a suggested move highlighted on the board. These moves may or may not be optimal and you are not required to follow their suggestion. Have fun playing!

Post-treatment information page

Thank you for playing. Now, please complete the post-game survey by clicking this link.

Sources Cited

Griggs, Brandon. 2011. "Why Computer Voices Are Mostly Female." *CNN*. Cable News Network. https://www.cnn.com/2011/10/21/tech/innovation/female-computer-voices/index.html.

Hesse, Monica. 2019. "Elizabeth Holmes's Weird, Possibly Fake Baritone Is Actually Her Least Baffling Quality." *Chicagotribune.com*. Chicago Tribune. https://www.chicagotribune.com/entertainment/tv/ct-ent-elizabeth-holmes-theranos-voice-20190321-story.html.

Mulaney, John. 2019. "Saturday Night Live." *Saturday Night Live*. NBC.