Speaker’s Facial Expressions Impact on Learning Outcomes

Group 11
Curt Williams
Klaijan Sinteppadon
Lisa Yeung
Marissa Ferrighetto
Rahila Sule

Executive Abstract

Online learning has become more prevalent because of the global disruption of the COVID-19 pandemic. Interested in how the magnitude of facial and head movements of the speaker impacts a listener's learning, we conducted a randomized control trial in which online participants, listeners, watched short five minute video lectures about A/B testing recorded by students. This study demonstrate that facial and head movements of the speaker in an instructional video on A/B testing does not increase or decrease understanding of the content in this population of subjects.

Introduction

Problem statement

Because of the COVID-19 pandemic, traditional models of teaching have been disrupted. Modes of classroom instruction have shifted from primarily in-person lectures to entirely online synchronous courses, online asynchronous courses, and hybrid instruction. Since the pandemic has accelerated online lectures and learning, we decided to explore how an instructor's expressiveness affects a student's learning outcomes. In particular, we look at whether the magnitude of an instructor's movements throughout an educational video affects the level of understanding of the material. Does an instructor who moves more keep a listener more engaged? Or does a greater magnitude of movement distract a listener from the content? Would movement of the instructor impact a student's understanding and performance? To examine these questions, we contextualized our study among the existing literature, designed the below randomized control trial, described in detail in the "Empirical Experimental Setting" section, and analyzed the results.

Literature review

Researchers have posed similar questions about how facial expressions impact learning outcomes for both human instructors and virtual pedagogical agents. In "The emotional design of an instructor: body gestures do not boost the effects of facial expressions in video lectures," the authors examined the combinations of body gestures and facial expressions, concluding that the optimal combination associated with the highest student learning outcomes is happy facial expressions with no gestures (Pi et al., 2022). In their paper on designing nonverbal communication for pedagogical agents, Baylor and Kim show results that suggest the use of facial expressions enhance learning when they are not paired with gestures. The presence of facial expressions could have a positive effect on learners' perceptions of and attitudes towards the material during attitudinal learning (Baylor and Kim, 2009).

A similar study found that instructors deem dramatic facial expressions both for themselves and "for the attainment of students' learning outcomes." (Butt and Iqbal, 2011). Additionally, in an effort to ameliorate obstacles related to distance learning, Maqableh, Alzyoud, and Zraqou study how different online platforms contribute to learner engagement (2022). They find that more interactive instruction results in better performing students, but this involves innovative teaching methods beyond simply interactive gestures. Our study builds on this prior work by determining whether the magnitude of an instructor's facial movements affect the learner's understanding of content related to experimental design techniques such as A/B testing.

Empirical Experimental Setting

Data Collection and Overview

For this experiment, roughly 20 groups of students created five-minute instructional videos with the goal to teach A/B testing to folks that might not be familiar with the basic principles of A/B testing. A subset of the videos were chosen for quality and accuracy in answering the learning objectives. Each video showed exactly one face, large enough for the listeners to see their facial expressions. In the context of this study, we refer to the people in videos teaching content as "speakers" and the experimental subjects as "listeners." In the experiment, listeners were asked to answer pre-video questions to assess their existing understanding of the content; they were then randomly assigned exactly one video to watch. After the video, listeners answered eight content-related questions, aiming to measure their understanding of the basic principles of A/B testing explained in the video, as well as demographic questions to better inform the study.

The data was collected in November and December 2022, and the experiment was conducted online. While the study population mainly consisted of a convenience sample based on student professional networks, families, and friends, students in the course represented a fairly diverse population, leading to a fairly diverse set of respondent subjects. The full sample comprises 255 respondents. The videos were labeled using Azure. Amongst all the 23 videos, only 19 were shown to the population. 1 of the 19 videos was erroneously labeled with the speaker's gender as both male and female, thus, we dropped the video to avoid interpretation issues in our analysis. Consequently our final sample size comprises 227 respondents.

We performed preliminary data analysis on the variables of interest and their distribution is as follows:

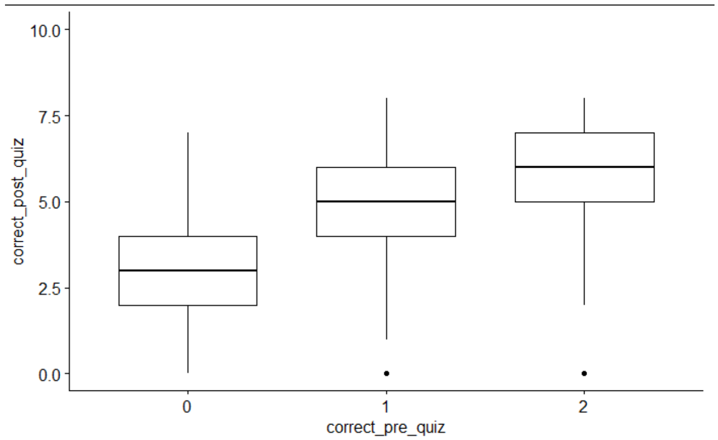


Figure 1. *Distribution of correct post quiz responses by number of correct pre quiz responses*

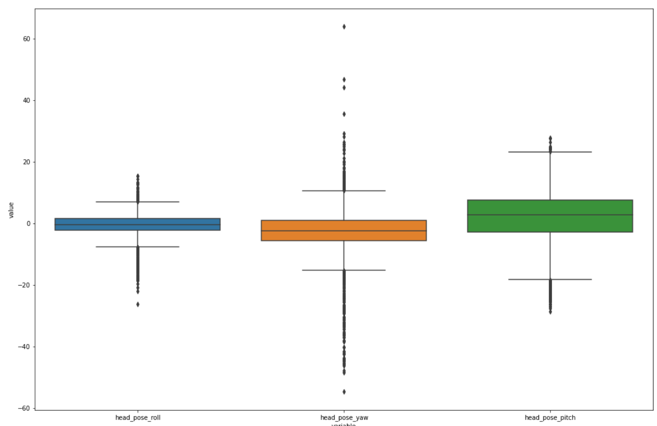


Figure 2. *Distribution of headpose variable movements*

The distribution of variables related to head pose shows high variations – which indicates that it may be a good candidate as a proxy for a speaker's movement, our question of interest. Samples with higher correct pre quiz have higher correct post quiz. The confounding relationship to our dependent variable suggests that we should include it as a control variable. In later sections, we explore setting up our regression analysis by considering any confounding relationship on dependent variables and include them as control variables and finding suitable candidates as independent variables based on collinearity analysis. The distributions of these variables are included in Appendix A.

Collinearity Analysis

After an in-depth exploration on the data available, we performed collinearity analysis on features that we are interested in before conducting regression analysis to ensure the robustness of the causal inference. This assessment is first conducted on different facial parts 1) nose 2) eyebrows 3) eyes and 4) headpose. All of the variables measuring movement of these body parts are assessed. The correlation analyses of other variables are found in Appendix B.

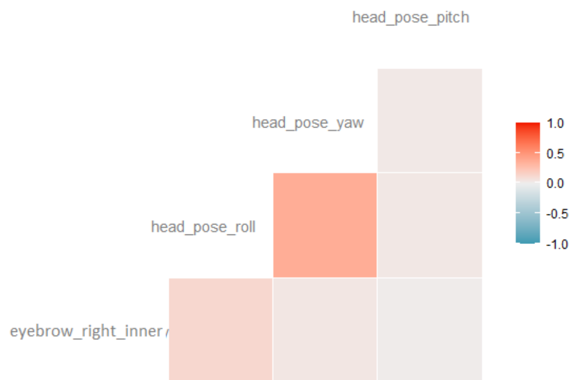


Figure 3. *Correlation analysis of variables used in "total movement" score*

From the correlation heatmaps, high collinearity is evident among the nose, eyes and eyebrows variables, showing a pattern that the left side and right side are highly correlated within these variables. On the other hand, the headpose variables are not highly correlated with each other, except head_pose_yaw and head_pose_roll which are moderately correlated. Due to the high correlation among the facial feature variables, many variables were dropped and only the low to moderately correlated variables were used in calculating "total movement" score.

Experimental Design

To examine the effect that a speaker's facial movement has on a listener's understanding of an educational video's content, we run a randomized controlled trial on the subjects by randomly assigning videos to subjects.

First, we computed a score for the speaker's facial movements. We used the variation of a speaker's facial positioning as a proxy. We normalized each variable related to a speaker's facial positioning. This included head positioning and facial feature positioning (mouth, eye, eyebrow, lips, nose). Next, across all videos, we computed the variation across captured frames grouped by video_id to get the total movement for each facial feature in each video. This variation showed how much a speaker moved their facial features over the duration of captured frames in the video. We then summed all facial feature variables to get a "total movement" score. After determining multicollinearity among many of these facial features, however, we recalculated the total movement score for features that were not collinear: *head roll, head pitch, head yaw, and inner right eyebrow.* To ensure that these features were still useful proxies for facial movement, we compared the scores from these features to the cumulative scores including all features. The distribution of videos ranging from high scores to low scores was consistent across both types of scores. Additionally, we ran a t-test for both types of scores and the results remained consistent. We then computed the median movement score. All videos that had a movement score above the median were considered "treatment" i.e. received videos with expressive facial movements. All videos that had a movement score below the median were considered "control" i.e. received videos with minimal facial movements.

The most important part of the experimental setting is randomly assigning the treatment to treatment and control groups. Since the videos themselves are randomly assigned, there was random assignment of videos with expressive facial movements and minimal facial movements. **This implies that all listeners had an equally random chance to receive a video with a large variation in facial positioning, or the treatment, and a video with a small variation in facial positioning, or the control.** First, we split the videos into two groups based on our measure of expressive versus minimal facial movements of the speaker. Survey respondents (listeners) that watched videos with expressive facial movements were assigned to the treatment group while listeners that watched videos with minimal facial movements were assigned to the control group.

Random assignment was especially important in this experimental setting because it allows us to infer causality between the movement in a video that the listener watches and the amount of answers the listener responds to correctly. Vitally, without random assignment of the video watched, we would not be able to infer causality – at most we could infer an association among the movement in a video and the outcome.

With this randomized controlled experiment, we aim to measure the causal effect of the magnitude of a speaker's expressiveness (i.e. facial and head movements) on a listener's understanding of A/B testing (as measured by their post-quiz results).

Importance of the Randomized Controlled Trial

Randomized Controlled Trials allow us to claim causality when examining the effect of a speaker's movement on the listener's understanding of the content. It guards against several issues. First is the counterfactual. To find the causal effect, we compare the treatment and control outcomes. However, a subject can only receive the treatment or the control, but not both. Thus, the outcome that we do not have for a subject is the counterfactual. For example, if a listener is assigned a video with a lot of movement, we observe their outcome given the movement but we cannot observe what the listener's outcome would have been if they had been assigned a video with very little movement because they were only assigned one video. This problem is called the *counterfactual*.

Next we'll discuss *selection*. It is tempting to compare the treatment and control outcomes across different subjects because we may have vast data for both subjects. In a setting where the treatment is not randomly assigned, this would be a mistake, even when the subjects are similar on factors that we observe. In that setting, the comparison will always have an error of some amount; the error is called selection. Selection measures how subjects may have been different to begin with. There may always be differences between people that we cannot measure that contribute to differences in outcomes, and thus affects our ability to say what causes the outcome. For example, if all the males in our population were assigned low movement videos and everyone else was assigned the high movement videos.

Selection introduces the issue of the *backdoor*. A backdoor, referred to as Z, represents variable(s) that affect the outcome of an experiment, independent of our causal variable of interest. In our context, Z would be a variable related to the number of post-video correct answers that is also related to the variation in movement in a given video. Random assignment of the treatment, however, negates the effect of the backdoor – if a listener receives a video with an equally random chance to have movement or not, no backdoor should exist such that $\boldsymbol{Cov(y, Z) = 0}$ and $\boldsymbol{Cov(x, Z) = 0}$.

To avoid these problems, we randomly assigned the treatment to subjects, as explained above. After random assignments, the two groups should, on average, be the same. This will balance all factors by ensuring that things unobserved across subjects that are assigned the speakers with high movement are similar to factors unobserved for subjects assigned speakers with low movement and therefore allow us to determine causal effect.

Regression

We use a simple OLS regression to obtain the causal effect. Our specification is detailed below:

$$y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{n} \gamma_j \theta_{ji} + \varepsilon_i$$

Where:

y_i	is the post-quiz number of correct answers of each subject, i
β_0	is the constant, i.e. the baseline of average correct answers when the observing subject belongs to the control group
β_1	is the estimated causal effect, i.e. the changes in average correct answers compare to the baseline when the observing subject belongs to the treatment group
T_i	is a binary variable for each subject, i, 1 for subjects in the treatment group and 0 for subjects in the control group
γ_j	are the coefficients for the corresponding control variable, j, i.e. the amount of change the control variable contributes to the correct answers
θ_{ji}	are the following control variables, j, included in the regression for each participant, i: Total number of correct answers pre quiz, and speaker's smile
ε_i	is a residual, i.e. the difference between the predicted value and the actual value

Control Variables

We included several control variables in this study that could impact the outcome. Most importantly, we controlled for participant's knowledge on A/B testing prior to watching the video as measured by *correct_pre_quiz*. We controlled for the speaker's average smile and gender; two factors that may have an impact on listener understanding beyond our causal variable of interest. We controlled for factors related to the listener, such as race, gender, level of education, and reported interest in the video. The speaker's gender and listener's gender interacting may have an effect on the outcome, so we also controlled for the interaction between these terms. After dropping variables that are statistically insignificant, we have *smile* and *correct_pre_quiz* as our control variables.

Results Obtained

Descriptive Statistics

There were 135 listeners in the control group, and 120 in the treatment group. The groups are characterized on the variable in focus as follows:

Table 1

	Male	**Female**	**Other**	**Unspecified**
Control	58	52	4	6
Treatment	60	43	0	4

Table 2

	Control				**Treatment**			
	Min	Med	Mean	Max	Min	Med	Mean	Max
Correct pre quiz	0.0	1.0	1.225	2.0	0.0	1.0	1.271	2.0

Videos are characterized as follows:

Table 3

Speaker's gender	**Male**	**Female**
Control	3	5

Treatment	5	4

Table 4

	Control				**Treatment**			
	Min	Med	Mean	Max	Min	Med	Mean	Max
Smile	0.00	0.06	0.17	0.59	0.00	0.03	0.08	0.26

Results

We found that the speaker's movement in a video does not have a significant effect on a listener's understanding of the basic principles of A/B testing, as measured by the number of post-quiz correct answers. We can say neither that a speaker's movement keeps listeners engaged, nor that a speaker's movement distracts from the content of the video. This is evident by the coefficient for *group* (i.e. being in the treatment group) being insignificant in our regression analysis (Table 5.). We note that the insignificant result does not necessarily mean that the treatment had no effect on the number of correct answers, it could be that the effect of the treatment was too small to detect with the sample size of the study.

The regression results are shown in *Table 5*.

Table 5: *Randomized Controlled Trial Regression Results*

```
=================================================
                    Dependent variable:
                 --------------------------
                     correct_post_quiz
---------------------------------------------
group                    0.114
                         (0.226)
smile                    1.255*
                         (0.643)
correct_pre_quiz         1.147***
                         (0.143)
Constant                 3.181***
                         (0.254)
---------------------------------------------
Observations             227
R2                       0.235
Adjusted R2              0.224
Residual Std. Error      1.635 (df = 223)
```

F Statistic	22.777*** (df = 3; 223)

==

Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 4. further illustrates the non-significant effect, displaying near identical distribution of correct answers by those who viewed videos with a lot of movement by the speaker and those who viewed videos with little movement by the speaker.

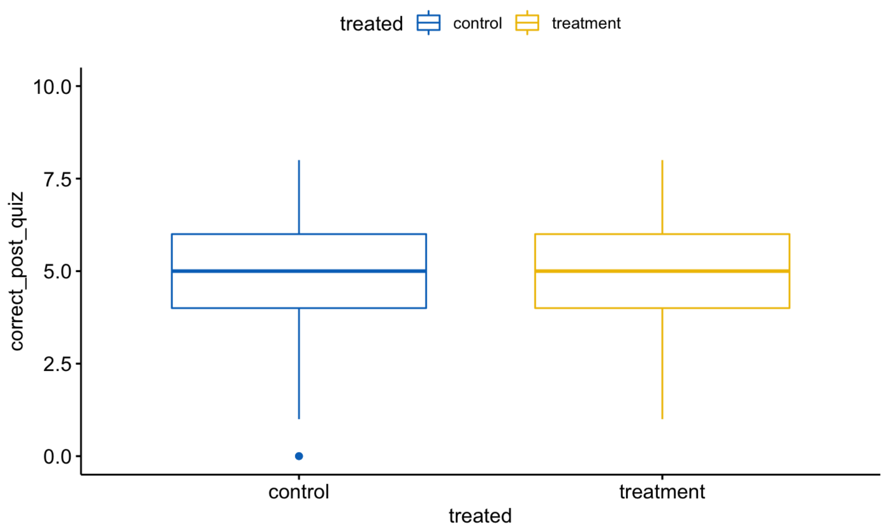


Figure 4. *Distribution of Correct Post-Quiz Answers by Treatment Group*

Additionally, we show the mean, standard deviation, and median of the correct answers in both types of quizzes for both groups below:

	Mean Correct Pre-Quiz	Standard Deviation	Median
Control	1.23	0.77	1
Treatment	1.27	0.75	1

	Mean Correct Post-Quiz	Standard Deviation	Median
Control	4.80	1.95	5
Treatment	4.85	1.76	5

There is not a significant difference between the control and treatment groups for post-quiz mean at all significance levels of 0.01, 0.05, and 0.1 based on p-value of 0.613.

We also checked for heterogeneous causal effects to see if the treatment affected different subpopulations of our sample differently. For example, we checked for whether the treatment affected different genders, races, education levels, and other factors to get significantly different proportions of correct answers in the post-video quiz. We did not find any evidence of heterogeneous causal effects for any factor, the results of this are displayed in Appendix C.

Conclusions

There is no effect between the magnitude of a speaker's facial movement in an instructional video and a listener's understanding of the content of that video in the experimental context outlined in this report. Despite controlling for factors such as gender, prior understanding of A/B testing, educational background, race, and interest in the material, the results of the randomized controlled trial were insignificant. These studies prove that facial movement in an instructional video on A/B testing for the subpopulation of our subjects does not increase or decrease viewer's understanding of the content.

However, these results do have implications for the world of online educational content that is growing in popularity. Sometimes, instructors are more worried about their appearance and how that affects students' learning than the content. While our results do not generalize to a broader population or other online learners, instructors of this content for passive learners may not need to be hyper-focused on their movements and can focus fully on the content.

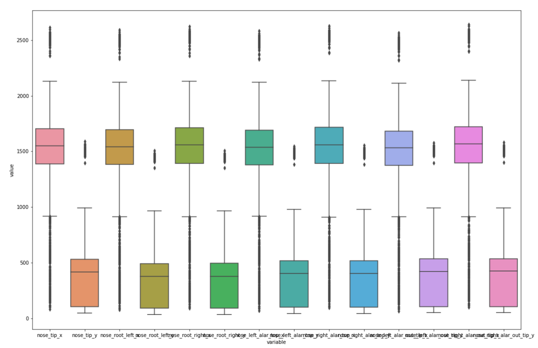
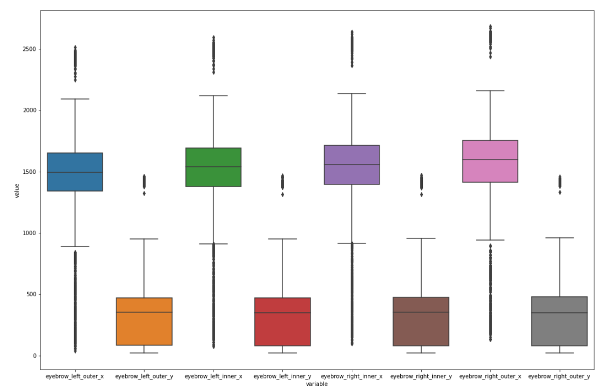
This study came with several limitations. Most glaringly was the lack of data. If more subjects had committed to this study, we may have seen more of a difference between the two groups. Additionally, the scope of the study was narrow in multiple ways. First, the range of subjects we encountered did not represent the diversity of online learners. Our sample was limited to those 18 and older and also limited to those within the network of students in the course. A wider population of subjects may result in different conclusions. In the future, it would be interesting to see if movement of a speaker affects understanding across different types of learners, such as those of different ages (from elementary age to senior citizens) or those of different cultural backgrounds.

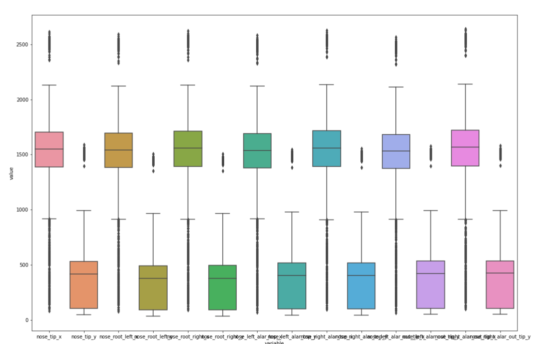
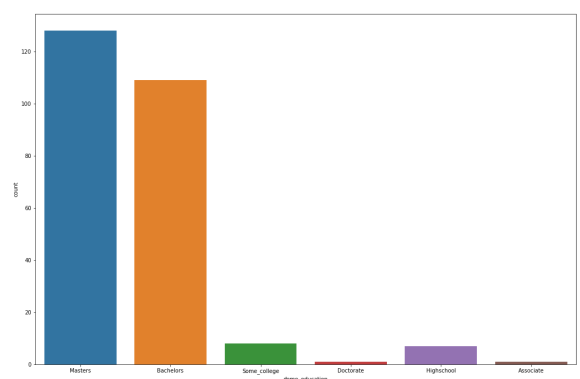
Second, the scope of subject material was limited to A/B testing. In the future, it may be of interest for researchers to see if the conclusions change with the content of the videos. For example, does more dramatic instructor movement make history-based, science-based, or literature-based video content more or less engaging? In other words, does the pedagogy change with subject matter?

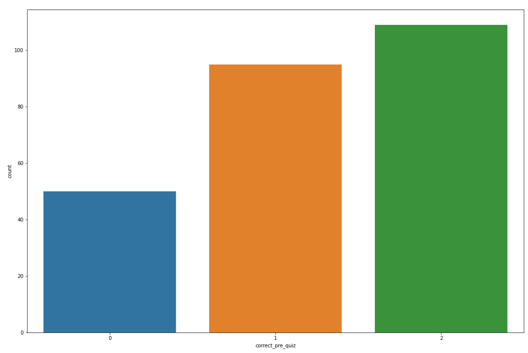
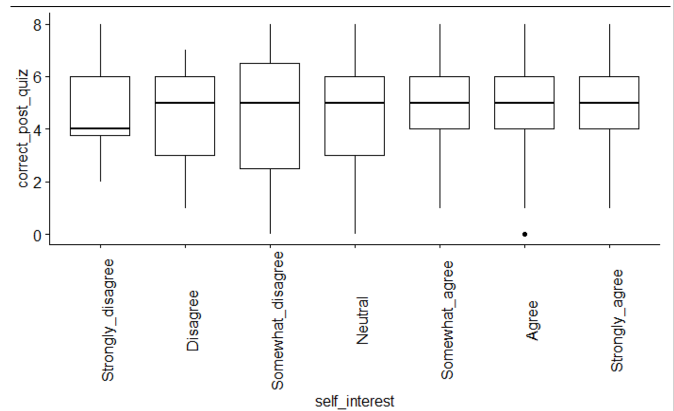
Finally, future studies could purposely vary the movement more dramatically. The sample of videos that were randomly assigned in this study were made without this experimental question in mind. Videos that more purposefully vary the instructor movement in the video may display more overt differences in a learner's understanding of the material.

Appendix A

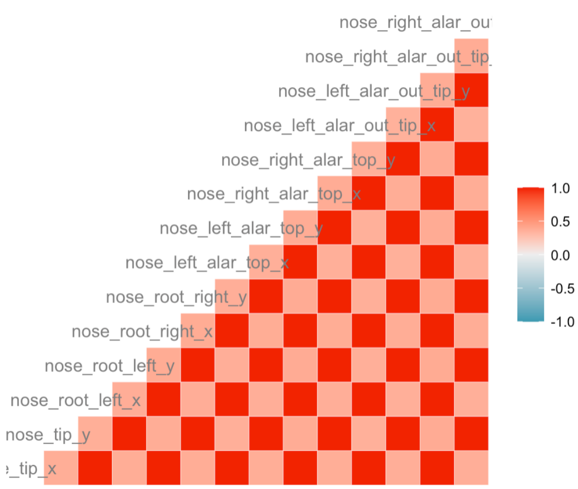
Pre-study descriptive analysis of factors of interest. This analysis helped us to determine which variables to use to split treatment and control and also helped us to brainstorm variables that would be useful to control for.

	
Distribution of variables related to eyes	Distribution of variables related to eyebrows

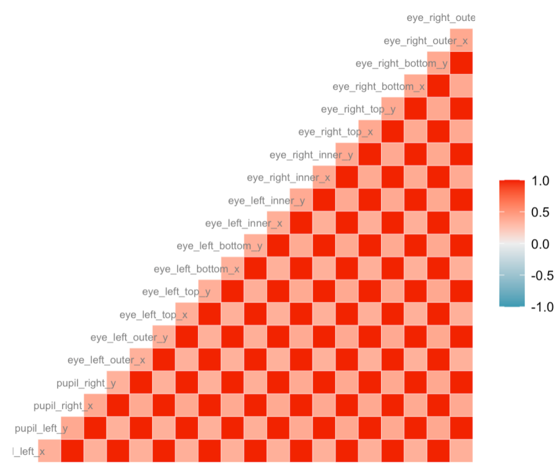
	
Distribution of variables related to nose	Distribution of participant's education level

	
Distribution of correct pre quiz	Relationship between a listener's reported self-interest and the number of questions they answered correctly in the post-video quiz

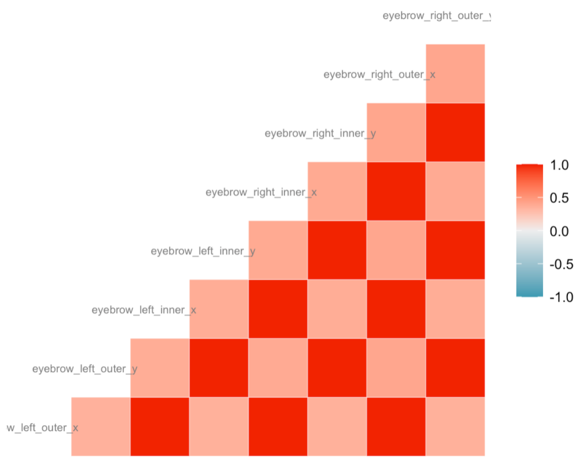
Appendix B



Correlation analysis of variables on nose



Correlation analysis of variables on eyes



Correlation analysis of variables on eyebrows

Appendix C

Test for heterogeneous effect with gender

	Dependent variable: correct_post_quiz
group	-0.045
	(0.328)
I(demo_gender)Female	-0.292
	(0.333)
I(demo_gender)Other	-3.883***
	(0.902)
I(demo_gender)Unspecified	-1.984**
	(0.795)
smile	1.198*
	(0.687)
group:I(demo_gender)Female	0.173
	(0.482)

group:I(demo_gender)Other	
group:I(demo_gender)Unspecified	-0.700
	(1.172)
Constant	4.951***
	(0.257)
Observations	227
R2	0.144
Adjusted R2	0.116
Residual Std. Error	1.745 (df = 219)
F Statistic	5.242*** (df = 7; 219)
Note:	*p<0.1; **p<0.05; ***p<0.01

Note: Controls used are: Participant's gender, participant's level of education, participant's race, total number of correct answers pre quiz, speaker's gender, speaker's smile, total duration of survey, participant reported interest in video

	Dependent variable:
	correct_post_quiz
group	-0.112
	(0.219)
smile	0.866
	(0.750)
speaker_gendermale	0.086
	(0.324)
survey_duration	0.00002
	(0.00002)
self_interestDisagree	-0.587
	(0.702)
self_interestSomewhat_disagree	0.382
	(0.607)
self_interestNeutral	0.889
	(0.553)

self_interestSomewhat_agree	1.186**
	(0.572)
self_interestAgree	0.749
	(0.540)
self_interestStrongly_agree	0.531
	(0.575)
demo_raceBlack	-0.700
	(0.946)
demo_raceAsian	-0.222
	(0.396)
demo_raceNative_American_Alaska	-1.167
	(1.619)
demo_raceNative_Hawaiian_PacificIslander	-2.078
	(2.281)
demo_raceNone	0.189
	(0.854)
demo_raceUnspecified	-1.460**
	(0.581)
demo_genderFemale	0.032
	(0.315)
demo_genderOther	-2.535*
	(1.358)
demo_genderUnspecified	-1.101
	(0.962)
demo_educationSome_college	0.547
	(0.998)
demo_educationBachelors	0.804
	(0.793)
demo_educationMasters	0.633
	(0.780)
demo_educationDoctorate	
correct_pre_quiz	0.971***
	(0.143)
speaker_gendermale:demo_genderFemale	-0.561
	(0.430)
speaker_gendermale:demo_genderOther	0.029
	(2.148)
speaker_gendermale:demo_genderUnspecified	0.826
	(1.147)
Constant	2.626**
	(1.054)

Observations	227
R2	0.402
Adjusted R2	0.324
Residual Std. Error	1.527 (df = 200)
F Statistic	5.164*** (df = 26; 200)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix D - References

Baylor, A. L., & Kim, S. (2009). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior*, *25*(2), 450-457.

Butt, M. N., & Iqbal, M. (2011). Teachers' Perception Regarding Facial Expressions as an Effective Teaching Tool. Contemporary Issues in Education Research, 4(2), 11-14.

Maqableh, W., Alzyoud, F. Y., & Zraqou, J. (2022). The use of facial expressions in measuring students' interaction with distance learning environments during the COVID-19 crisis. Visual Informatics.

Pi, Z., Liu, R., Ling, H., Zhang, X., Wang, S., & Li, X. (2022). The emotional design of an instructor: body gestures do not boost the effects of facial expressions in video lectures. Interactive Learning Environments, 1-20.