

Study of Engagement in Online Learning

12/10/2020 - Final Report S4 & S5

Prof. Alex Hughes

Bryan Auyeung, Alexandra Savelieva, Dana Kaban, Vasanth Ramani, Casey King

Abstract

Recently, more professional and personal social activities have been moving into digital space - the general global trend that has been dramatically accelerated by COVID-19 restrictions. Education has been impacted by this shift very substantially. But while online classes impose a different level of complexity on instructors, there are new opportunities to improve studies on student engagement. In this work, we rethink how engagement can be measured in controlled and naturally occurring experiments by proposing a new research framework that scales traditional methods using surveys by collecting thousands of signals with computer vision and DL techniques. Our engagement detection algorithm is bootstrapped with a custom labeled dataset that was created for training a custom deep neural network. The proposed solution is tested on a voluntary study of Berkeley MIDS students with the focus of evaluating impacts of topic-related discussions on engagements in online classes over Zoom.

Keywords

Learning engagement, controlled experiments, causality, research design, statistical methods, controlled experiments, computer vision AI, NLP, deep learning, data engineering, multimodal

Introduction

Engagement is a key indicator in understanding the effectiveness of education informational systems. It is defined and measured differently across various applications such as search engines, sales teams, and mobile apps. At the intersection of human-technology interactions and academia, engagement plays a critical role in the learning process and has important bearings on learning outcomes.

As MOOCs and personal development platforms are growing in popularity, it is increasingly important to understand learner engagement in an online environment. The proliferation and challenges of online education also highlighted by the migration of classes to webcam recorded lectures due to the COVID-19 epidemic. As educators turn to video platforms and e-learning tools, the question of learning effectiveness persists: Are students readily understanding schoolwork and class concepts in spite of the distractions, learning curves, and accountability measures associated with an online environment? [1] Despite the proliferation of online platforms, the in-person connection in classrooms is often heralded as the linchpin for academic success.[2] According to the National Center on Safe Supporting Learning Environments, engagement is defined as "*strong relationships between students, teachers, families, and schools, and strong connections between schools and the broader community.*" In our study, we

analyze the level of engagement in students and focus on engagement measures of student to student interactions within online classes. We believe the results in our experiment can give meaningful insights to learning in all environments.

Prior work

Traditionally, measurements of engagement are taken through observation or surveys [L. Schultz & J. Sharp, 2007]. This presents a number of challenges involving systemic bias and interference from the research observer. For studies involving surveys, there are high risks related to low student participation rates and inaccurate responses. Both methods present difficult scenarios for obtaining results with significant statistical power.

More recently, research in the automatic evaluation of visual signals has been applied to engagement in the works of [O. Mohamad Nezami et.al., 2019; P. Goldberg et.al., 2019; J. Whitehill et.al., 2014]. By using deep neural networks, these authors have shown the power of interpreting engagement in still pictures of high school and college students recorded completing learning based tasks. In the Nezami et.al. research, high school students were instructed to play a computer game while their faces were recorded and analyzed with a VGG neural network model. College student participants in the Goldberg experiment were tested on topics of the lecture session and monitored with cameras throughout the duration of the class. Linear and SVM regressions were used on hand-labeled data points corresponding to facial features captured on video.

These efforts have set ground to the techniques and tools we used to analyze engagement in an online learning environment. In addition to the visual inputs that were used in prior work, our intent is to use conversational signals from voice transcripts and messages from classroom chat channels to converge on identifying engagement in student subjects. We want to extract, process, and estimate naturally occurring expressions of engagement in the form of facial expressions, poses and speech for the purpose of evaluating treatment effects of pedagogical experiments.

Our motivation is twofold: first, to inform theoretical understanding of human behavior in learning situations; second, to enable the development of computerized learning companions that could provide effective personalized assistance online. This work is tackling the problem set in [A.Savelieva, 2019a] and [A.Savelieva et.al., 2019b] to apply scientific analysis methods to automatically evaluate the efficiency of social interactions in different contexts.

Research Question

For the proof of concept, we chose to focus on a very specific question: *What is the impact of topic-related discussions via private chat on student engagement in online classes?*

Hypothesis

The null-hypothesis for our experiment is that experimental interference has no impact on student engagements in online classes. The definitions of components of this statement are as follows:

- *Interference*: Private IM on the lecture topic
- *Online classes*: Zoom web recorded sessions
- *Students*: Consenting students from UC Berkeley MIDS program
- *Engagements*: Speech, chat messages and observable face & body signals

We intentionally avoid making positive or negative assumptions regarding changes in engagement due experimental interference in order to analyze all possible results. Asking students about course material throughout a lecture can stimulate engagement among individuals who were not previously engaged or reinforces attentive behavior for those who are already engaged in class. It is also possible to see some students get distracted and change their focus from the lecture material to become more socially interactive, which could lead to reduced engagement. In our study, we intend to address this question by assessing the treatment effects at the individual and classroom level.

Insights from this study can be instrumental in such areas as development of computerized learning companions or optimization of configurations of educational platforms interaction tools.

Ethical Considerations

Ethics is broadly defined as “the correct rules of conduct necessary when carrying out research.” [Belmont Report, 1979]. We recognize that as researchers we have not only an obligation but a duty to protect research participants from harm. In our case, it was a bit of a challenge. We were worried that our consent form would compromise our experiment, and would signal to the participants that they were about to be studied. But we decided to set aside our concerns, recognizing that our obligation is to respect the rights and dignity of research participants. To that end, all students received consent forms. If this were any other study conducted for anything other than a class project, we would have to have sought Institutional Review Board Approval from Berkeley. The educational purposes of this student project offers an exemption to this mechanism. However, if we want to attempt to publish our results, we would, in fact, need to seek approval from the IRB. As a matter of fact, many of the top journals will refuse to publish research without IRB approval.[Klitzman, 2011]

Power Calculation

Prior to conducting pilot studies, we hypothesized sample size scenarios where a 10% change in engagement treatment effect was statistically significant compared to a 50% engaged control group. With a standard deviation of 20%, the sample size required for 80% power was 63 subjects in control and 63 in treatment. We attempted to recruit students across all sections of 241 and 266 for the following experiments and continued to track the requirements for a high powered experiment by using data from one of our successful pilot studies. An updated calculation [Figure 1] for attaining 80% chance of getting a p-value of less than 5% in our statistical test showed that we would need approximately 300 consenting participants with the treatment we saw in Pilot 1 (w241, section 5).

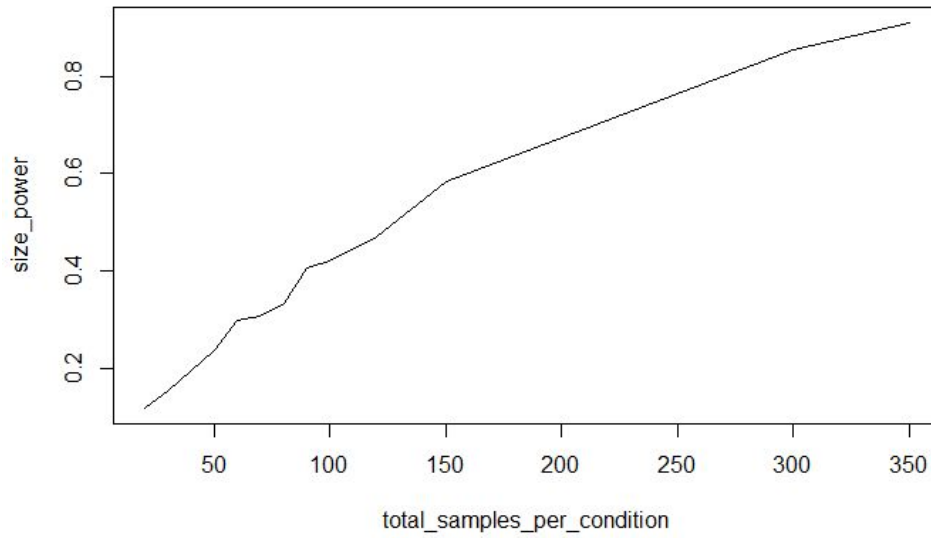


Figure 1: Pre-Experiment Power Calculation using Pilot Study 1 (section 5)

Experiment Design

Each class section was assigned two randomized time blocks for experimentation. All consenting subjects were randomly categorized into 2 subgroups. During the first experimental study, the first subgroup of students was treated and the second subgroup of students fell in the control group. After the results from the treatment effect were collected, the second experimental study was conducted at the second randomized time slot. The two subgroups switched roles during the second study, following a difference in differences model. The ROXO grammar for the experiment is shown in the Table 1 below. We used the pretest-posttest control group design and measured the baseline before the treatment and measured the outcome after the treatment as seen here.

Table 1. ROXO design

One Session	Subjects (randomly distributed between subgroups)	Treatment (Single-blinded)
Study 1	Subgroup 1	R O X O
(randomly picked time)	Subgroup 2	R O -- O
Study 2	Subgroup 1	R O -- O
(randomly picked time)	Subgroup 2	R O X O

Randomization

We initially distributed consent forms to 58 students of the 63 students that could be potentially targeted for this experiment. 37 out of the 58 students consented and were monitored throughout our experiments. The flow document for two pilot sections is shown in the diagram below. The sections are taken from w241: Experiments and Causality and w266: NLP with Deep Learning.

The treatment times are randomized for the two studies carried out in each section. Out of 6 participants in the first section, half were assigned for treatment at 4:30PM. The remaining half becomes the control group for the 4:30PM time slot and becomes the treatment group for the second study at 5:00PM.

In the second pilot, 3 subjects out of 7 were assigned to treatment during the first study which was conducted at 4:50PM. The remaining 4 subjects were treated during the second study. Subjects were monitored 5 minutes prior to the treatment until 7 minutes after the treatment to capture any delayed treatment effects.

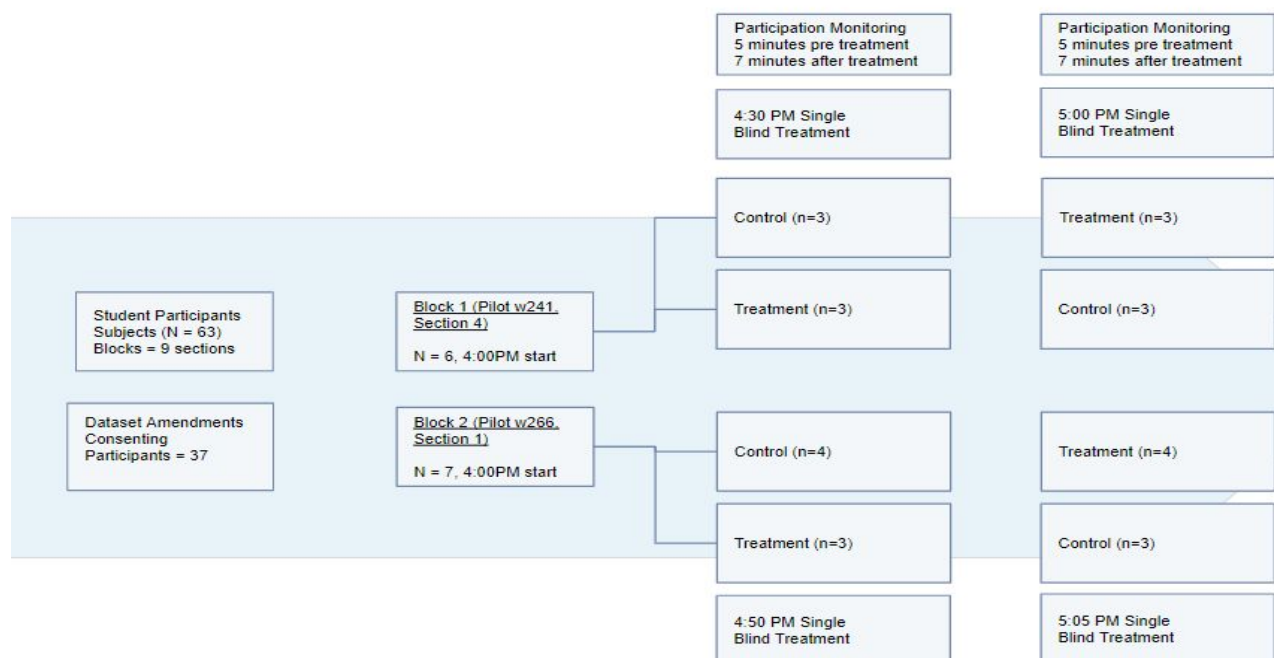


Figure 2: Experimental Process Flow

Proposed Treatment

All the consented subjects were identified at the start of the class. Subjects who were absent from the class were omitted from the random assignment process. Treatment questions for a specific class were prepared in advance of class and adjusted for each section so that no question was repeated. A custom algorithm generated random times and assigned individual researchers to execute the treatment for each subject in a section. In [Figure 3], we can see an example of the randomized treatment plan for a single section. All subject names are hidden to protect student privacy.

* Experimenter: Dana Kaban			
	date	subject	question experimenter
14	2020-10-29 19:00:00	[REDACTED]	what are we discussing? I missed it. Dana Kaban
* Experimenter: Casey King			
	date	subject	question experimenter
3	2020-10-29 18:40:00	[REDACTED]	what are we discussing? I missed it. Casey King
8	2020-10-29 19:00:00	[REDACTED]	what are we discussing? I missed it. Casey King
* Experimenter: Vasanth Ramani			
	date	subject	question experimenter
1	2020-10-29 18:40:00	[REDACTED]	what are we discussing? I missed it. Vasanth Ramani
2	2020-10-29 18:40:00	[REDACTED]	what are we discussing? I missed it. Vasanth Ramani
5	2020-10-29 18:40:00	[REDACTED]	what are we discussing? I missed it. Vasanth Ramani
12	2020-10-29 19:00:00	[REDACTED]	what are we discussing? I missed it. Vasanth Ramani
15	2020-10-29 19:00:00	[REDACTED]	what are we discussing? I missed it. Vasanth Ramani
* Experimenter: Alexandra Savelieva			
	date	subject	question experimenter

Figure 3: Assigned treatment

All treatment messages used in the experiment are detailed in [Table 2]. To avoid priming and repeating the same treatment questions, we created this group of treatments that fall under the topic of discussion for that class section.

Table 2. Actual treatment

Section	Treatment
Pilot Sec4 10292020	Sorry what were we discussing? I missed it
Pilot Sec5 10292020	What are we discussing? I missed it..., what are we discussing? I missed it.
Study Sec1 11172020	What were we discussing? I missed it
Study Sec2 11182020	What are we discussing? I missed it.
Study Sec3 11182020	hey <SUBJECTNAME>, what are we discussing right now? I missed it.
Study Sec5 11052020	Hey <SUBJECTNAME> did you get a handle on the conditions under which it's ok to use naturally occurring experiments? i'm still not sure i understand...
Study Sec5 11192020	What are we talking about? I missed it., Hey <SUBJECTNAME>, do you know what were we discussing? I missed it :/
Study Sec5 12032020	I just went to grab a snack. What paper are people talking about now?, was there a paper on this? or just the async info on it? I'm a little behind, I just went to grab a snack. What was the paper you sent?
Study w266 11032020	Hi. What are we discussing. I missed it.

Actual Treatment experience

Treatment was administered using the Zoom private chat messaging system. As seen in [Figure 3], the two studies carried out within the section are shown in blue (*Study 1*) and green (*Study 2*). Subjects were sent a treatment message privately at the proposed treatment time. The illustrated case shows the experimenter messaging “what are we discussing? I missed it” to subjects. Researchers recorded confirmation from subjects who acknowledged the message

and “accepted” the treatment. We have seen a spillover effect highlighted here in study 1 as one of the students sent the reply publicly in 1 section which could have alerted other students.

```

18:40:01 From Vasanth Ramani to SUBJECT(Privately) : what are we discussing? I missed it.
18:40:10 From Vasanth Ramani to SUBJECT(Privately) : what are we discussing? I missed it.
18:40:26 From SUBJECT to Vasanth Ramani(Privately) : #5 in ps4
18:40:32 From Vasanth Ramani to SUBJECT(Privately) : what are we discussing? I missed it.
18:40:33 From SUBJECT : SUBJECT is saying PS4 problem 5 is crashing on his side
18:40:58 From SUBJECT : |sorry, meant to reply privately to Vasanth :)|
18:41:42 From Vasanth Ramani to SUBJECT(Privately) : ok
18:41:50 From Vasanth Ramani to SUBJECT(Privately) : ok
18:42:33 From SUBJECT to Vasanth Ramani(Privately) : sorry for sending it to the wrong audience
18:42:59 From Vasanth Ramani to SUBJECT(Privately) : np man
18:43:18 From SUBJECT to Vasanth Ramani(Privately) : :)
19:00:12 From Vasanth Ramani to SUBJECT(Privately) : what are we discussing? I missed it.
19:00:24 From Vasanth Ramani to SUBJECT(Privately) : what are we discussing? I missed it.

```

Figure 4: Administration example

Data Collection Pipeline

At a high level, our experiment depends on signals collected from videos, speech and chat modalities. Image frames, transcripts and chat messages were processed to create our experiment dataset. The statistical methods and analysis is computed on top of this full dataset.

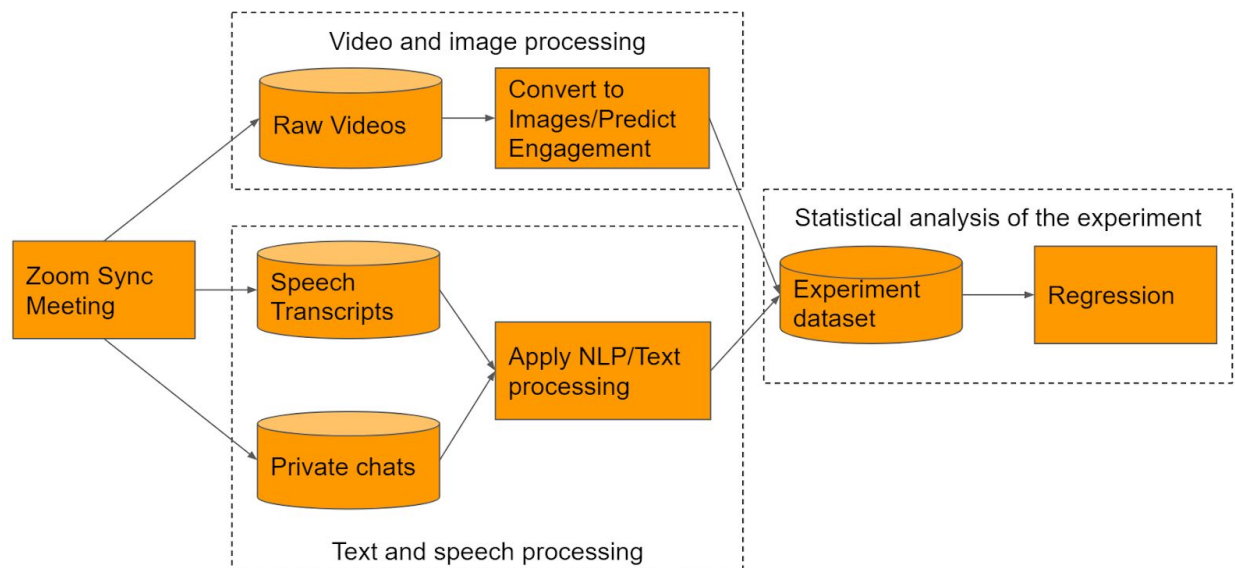


Figure 5: High-level pipeline design

Video and Image Processing Steps

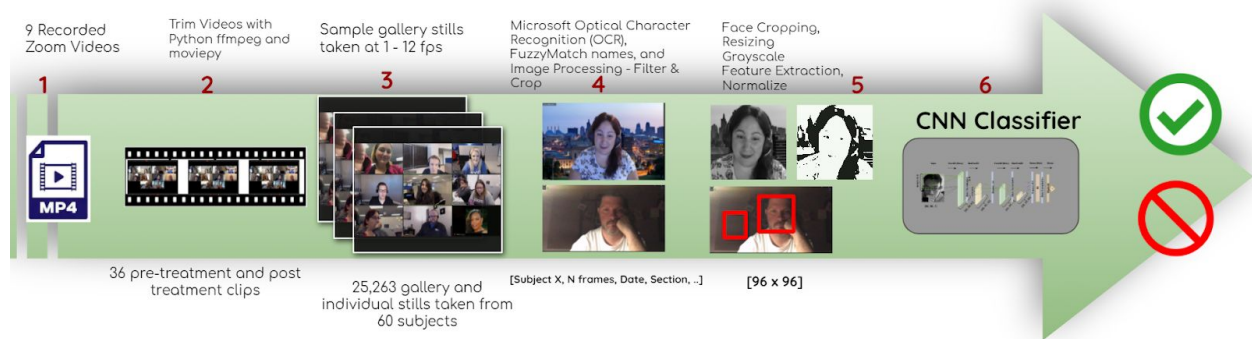


Figure 6: Pipeline design

Step 1: The input to this pipeline is our schedule of experiments. The output is an mp4 file of the class session recording, downloaded from zoom (or potentially any platform that allows to record videos, such as Teams).

Step 2: The input for step 2 is the treatment assignment file. The mp4 file from Step 1 is manually processed to extract Zoom gallery frames relevant to our experiments. We black out faces of individuals who have not been cleared with consented. The date and time of our experiment, section ID, and metadata regarding treatment administration is encoded in the file name. *Example:* `Sec5_12032020_730_pretreat.mp4` is the mp4 video recording for the experiment in Section 5 of w241 on the 3rd of December, pretreatment of the study conducted at 7.30PM.

Step 3: Files that were generated at Step 2 flow into a Jupyter Notebook that splits the gallery view into frames (at a rate 1 frame per second) and dumps corresponding JPG images into a folder named after the file. Each image encodes approximated time corresponding to the frame in PDT for ordering purposes, e.g. `sec5_frame_03_12_2020_19_30_02.JPG` for 7:30:02PM frame.

Step 4. With the help of OCR (Microsoft Optical Character Recognition), we extract the names of participants from each gallery view image. This API provides coordinates of the text blocks, from which we inferred the coordinates and corresponding sizes of the rectangles containing each subject's screen in the gallery view. Two types of errors are possible at this stage: some text blocks may not be detected (the issue is mitigated by iteratively applying picture cropping and name matching heuristics to help the API detect the text (e.g. "Dana Kabm" instead of "Dana Kaban" - mitigation by fuzzy string matching to the list of subjects and picking the closest match). In the cases where mitigation is not possible, two outcomes may happen: we fail to extract a frame for the person or we extract the wrong image. Output images are reviewed and erroneous images are dropped by researchers.

Step 5. Images from Step 4 are normalized (faces detected, cropped to the same size and converted to grayscale) and features for engagement prediction are extracted with the help of Python graphical packages.

Step 6. Images for each subject are fed into a deep neural network trained on pre-labeled data that produces a binary label (“disengaged” or “engaged”) for each image.

Text and Speech Processing

Text Processing

Private zoom chat messages were administered to students as prescribed by the proposed treatment plan. A history of all time-stamped chats were collected by the researchers at the end of each class. This data was preprocessed to separate researcher chats from relevant student responses. All responses are labeled according to time sent in order to join with video and speech data. To calculate the total treatment effect, we assign post-treatment scores to subsequent speech or video signals to measure the student’s engagement.

Speech Processing

Speech transcripts for each section were downloaded from the zoom recordings. The downloaded transcripts were cleaned to remove user avatars, users who have not consented, researchers and other omitted subjects. The recorded speech time was merged with the class start time and converted to the pacific standard time. Exact treatment/control time was removed from the private chats for both studies and are trimmed to capture the monitored time.

Engagement Recognition through Image Analysis

To determine changes in levels of engagement in student faces, we constructed a new dataset called the Student Engagement Dataset. The purpose of this dataset is for fine tuning a deep learning engagement model to recognize and determine engagement in MIDS students. The data is used solely for research purposes in training the model and is unavailable for public use. A total of 6 video recordings across 4 MIDS courses were used to create the images. These videos are provided to students within the Berkeley MIDS program. Utilizing the gallery view functionality within Zoom Web Recordings, the student samples were collected from the recorded videos at a fixed rate of 12 fps in random time frames spanning the length of the class session. A total 1,333 frames of student screens cropped from recorded videos of the virtual classroom were collected, giving us a representative distribution of facial expressions for subjects across different class sections and within individual classes. Images are normalized to 310x170 pixels and student names are blacked out from images for student anonymity.

Student Engagement Dataset

In order to fine tune the engagement model, we created a custom annotation software enabling annotators to independently label 1,333 sample images. This labeling process was adopted from techniques used by Nezami[2] and inspired by Aslan et.al.[13] Each sample is annotated by at least 2 annotators from our research group. Prior to labeling, each annotator is briefed with the behavioral and emotional dimensions of engagement as shown below.


Behavioral Dimension:

- **On-task:** The student is looking towards the screen or looking down to the keyboard below the screen.
- **Off-task:** The student is looking everywhere else or eyes completely closed, or head turned away.
- **Can't Decide:** If you cannot decide on the behavioral state.

Emotional Dimension:

- **Satisfied:** If the student is not having any emotional problems during the learning task. This can include all positive states of the student from being neutral to being excited during the learning task.
- **Confused:** If the student is getting confused during the learning task. In some cases, this state might include some other negative states such as frustration.
- **Bored:** If the student is feeling bored during the learning task.
- **Can't Decide:** If you cannot decide on the emotional state.

606. Zoom Picture Scoring



1. Select the person's behavioral appearance:

☐ On Task

☐ Off Task

☐ Can't decide

2. Select the person's emotional appearance:

☐ Satisfied

☐ Confused

☐ Bored

☐ Can't decide

On Task: The student is looking towards the screen or looking down to the keyboard below the screen.

Off Task: The student is looking everywhere else or eyes completely closed, or head turned away.

Can't Decide: If you cannot decide on the behavioral state.

Satisfied: The student is not having any emotional problems during the learning task. This can include all positive states of the student from being neutral to being excited during the learning task.

Confused: The student is getting confused during the learning task. In some cases, this state might include some other negative states such as frustration.

Bored: If the student is feeling bored during the learning task.

Can't Decide: If you cannot decide on the behavioral state.

Submit

Figure 7: Annotation Software

Images are first randomly selected from the Student Engagement Dataset and presented to the annotator for labeling. Given these behavioral and emotional dimensions, annotators labeled faces according to their understanding and interpretation of the image. A sample of the labeling software is shown in [Figure 7]. The labels are stored in a MongoDB database and become queried by the CNN model. In cases where an image receives more than two 'Can't Decide' labels, the image is removed from the training set. With this approach, the Student Engagement Dataset has 1,025 engagement and 308 disengagement photos.

Agreement

Many experiments rely on multiple people to label data. Due to the inevitable variability among human perspectives, significant differences in perception can lead to study error. In an effort to create a well-designed research study, one must consider and quantify this difference. This quantification of agreement is referred to as “interrater reliability.” Jacob Cohen recognized the problem with this approach—it did not include any consideration of agreement between raters that could be ascribed simply to chance. Cohen’s kappa [16] is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where:

P_o = the relative observed agreement among raters.

P_e = the hypothetical probability of chance agreement

We used an extension of Cohen’s Kappa known as Fleiss’s kappa to get an interrater reliability score of 0.413 across 3 researchers. This is consistent with prior work in automated engagement detection[2].

Engagement Recognition with Deep Learning

We are leveraging deep learning (DL) models' ability to learn hierarchical structures from low-to-high level feature representations to train an engagement classifier. We designed our architecture with the size of our dataset in mind (1,333 samples). Aiming to limit the model's capacity to learn and overfit on the train set while at the same time maximizing the features representations capabilities, we built a low complexity CNN model with ~1.57M trainable parameters.

The employed CNN model consists of two convolutional layers, a fully connected layer (50 units), and a final dense layer. The convolution layers have 32 and 64 filters with kernels of (5,5) and (3,3), respectively. (See figure below.)

Our DL classifier was trained and evaluated on hand-labeled data that was split using stratified sampling into a train (749 samples), validation (334 samples), and test sets (250 samples). The classifier achieves 79% and 74.4% accuracy on the validation and test sets, respectively.

We acknowledged various factors which could have introduced bias and hindered the model's performance. In hand-labeling the training data, we were given guidelines with room for interpretation which likely introduced selection bias. The data included randomized frames of the same set of students with slight positional variations and a significantly smaller number of disengaged images, resulting in an imbalanced dataset and potentially leading the model to favor the ‘engaged’ class label. To reduce these negative impacts on the model performance, we used feature contours and included a high dropout rate in the model.

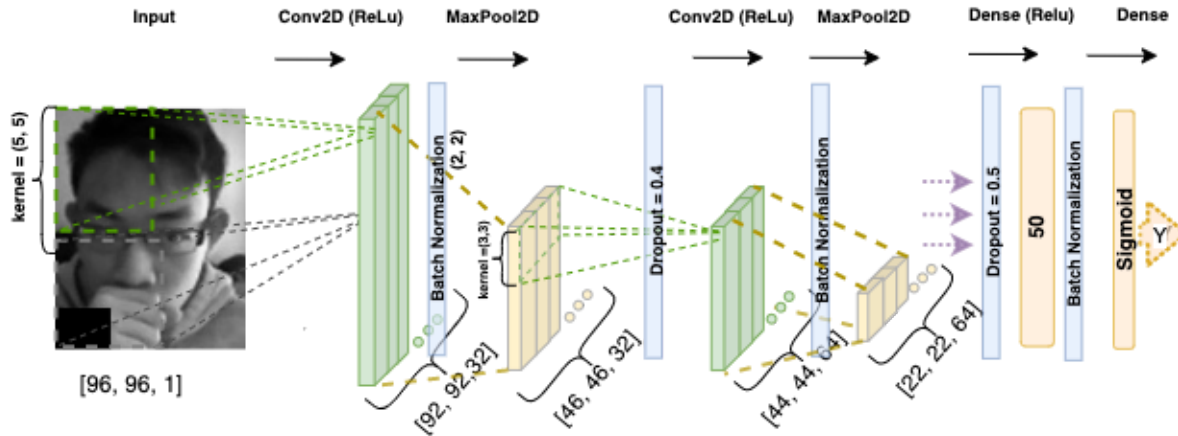


Figure 8: CNN model architecture used for engagement detection

Measurement units

The engagement data we collected can be broken down into audio and visual modalities. We observe approximately 470 frames per student in the combined pre and post-treatment when analyzing students' facial features and expressions in the class. In the 5-minute pre-treatment window, we consider the total number of frames collected as the total possible number of times an individual can be visually engaged. The same logic applies to the 7-minute post-treatment window.

Using our trained CNN model, we classify the pre-treatment and post-treatment frames as engaged or disengaged. The total number of frames labeled as engaged serves as the pre and post frames' score for the visual channel.

We count the combination of a single chat communication (textual modality) and a video transcript line (audio modality) as one engaged frame. We then add the text channel engagement score to the total number of frames and the visual engagement scores, giving a slightly higher weight to the transcript channel engaged unit in the final score calculation. We do this because speaking out in class generally implies engagement in the class.

Since the number of pretreatment frames is not equal to the number of post-treatment frames, we divide each of the scores by the total number of frames (# of engaged + # of disengaged frames) in the relevant set of frames. The resulting value represents the percentage of engagement during the relevant period (pre/post). As a final step, we compute the difference between the post-treatment and pre-treatment percentage scores.

Results

Data Quality

In our pilot analysis, we experienced treatment administration errors within Pilot 3, leading us to remove the pilot from our overall study. The professor in w266 disabled the private chat medium used for our treatment administration and created a Zoom breakout room during the initial

window of treatment. Our experimenter administered a delayed treatment using Slack (other messaging software widely used by students in the Masters program), which overlapped with the window of treatment for the other block of students in the same class. These 8 students were removed from our analysis due to the treatment spillover and consequent priming effect it had on the remaining students treated in the same class.

In our experiment, we observed situations with minor spillover and novelty effects. A student in one section responded publicly to a private treatment question, making their response visible to treatment and control groups. Given the nature of our blocked experimental design, we believe that the student's public response had a minor influence on other students and remains innocuous to the remaining block in the study section. If other students in the treatment group were engaged after witnessing the comments, they are likely to remain engaged or reiterate their response to treatment in the chat. Since the public response means little to the control group without the treatment context, the overall engagement effect would be miniscule across the 7 minute window in which we recorded data.

We also considered the novelty effects of having a substitute professor in a class section which could influence the initial engagement and attention in students during class. However, blocking by section constrains such effects to the section being studied. We also believe that the lasting effect from novelty exposure to a new teacher is minimal and would be present in the beginning of class, which is outside the window of the planned treatment.

We define noncompliance as completing our consent form but not acknowledging chat treatment messages over Zoom. In total, there were 27 compliers. We were unable to collect data on students who turned off their zoom cameras and didn't respond to messages. Students who dropped off of zoom during class or temporarily disconnected were considered attrited. Our measurement of engagement is contingent on recognizing facial images and responses to treatment. A total of 8 individuals attrited across control and treatment groups.

Exploratory Data Analysis: Study of group-level effects

If we look at the breakdown of data points [Figure 10] aggregated by subject for each administered treatment (or no-treatment, in case the subject is in the control group), we collected more samples in Section 5 and obtained higher sample variability. The number of treatments across control and treatment is well balanced in all sections except for section 4 due to attrition.

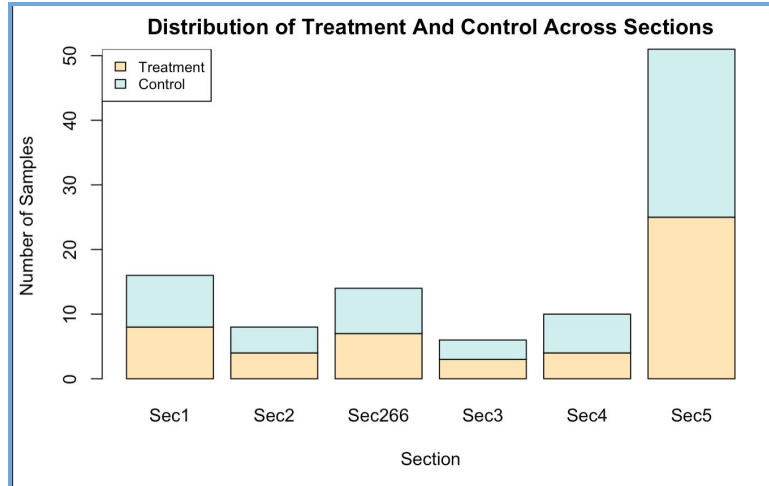


Figure 10: Distribution of Treatment and Control Across Sections

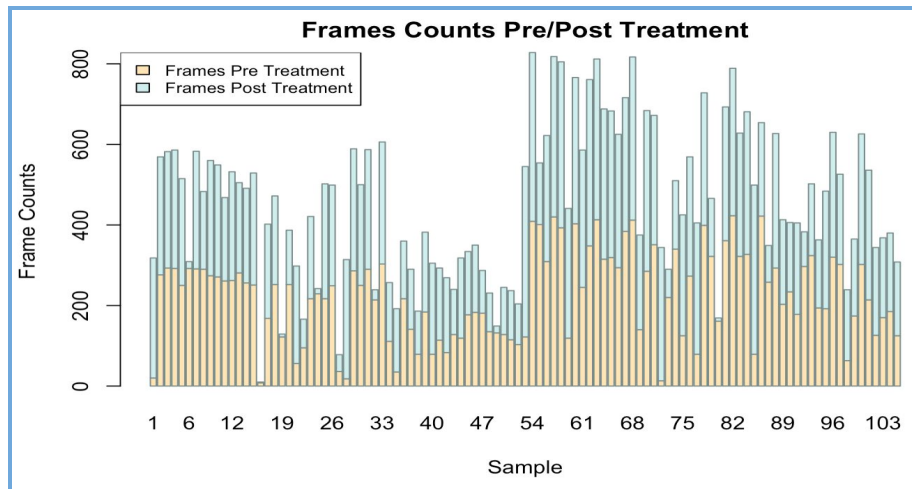


Figure 11. Frames counts pre/post treatment

Statistical Models

In the statistical analysis of our Randomized Complete Block Design Experiment, we aimed to estimate the combined weighted treatment effect across our treatment blocks [Tables 3, 4, 5]. In a total of four models, we gradually added blocks to the models such that the last is the complete model and includes the full set of blocks we used for randomization (date, section, study). We test these models on both Single Channel (Video) data and Dual Channel (Video and Audio) data.

Examining the coefficients of the results, we do not observe a significant estimate of the treatment effect. Utilizing the difference-in-differences model by taking differences in engagement pre and post-treatment, we aim to eliminate the variability in the treatment conditions that might bias our estimates.

However, we observe significance in the difference in treatment effect between blocks. On the surface, this suggests that some blocks were more impacted than others by our treatment. For example, in some sections, the class material could be more interesting one week compared to material in other weeks. Our treatment could reinforce engagement in students who are already attentive in class due to interest in the discussed material, but have no effect otherwise. However, there might be other factors contributing to this significant difference: our CNN classifier, which was our primary score measuring tool, produced an accuracy of 74.4% on the test data. Therefore, the model's biases might lead to over/under misclassifications of some sections frames resulting in inaccurate scores. Additionally, the treatment was administered by different people on certain days and sections. Subjects' reactions to treatment might differ based on their level of familiarity and perception of the treatment administrators, which could violate our excludability assumption.

We expected to see a decrease in our robust standard errors as we added additional blocks to the model; however, we observe the opposite. This might be due to random chance or due to differences between sections and a higher penalty to fit the regression line. As our estimated treatment effect is not significant, we are not concerned with this increase in the robust standard errors.

Table 3: Single Channel Models (Video Frames)

Single Channel (Video Frames) Models				
	Treatment	Sec (B)	Sec+Date (B)	Sec+Date+Study (B)
Treatment	-2.31 (1.95)	-2.08 (2.02)	-2.06 (2.07)	-1.78 (2.25)
Section 2		-0.82 (7.07)		
Section 3		2.47 (2.73)		
Section 4		-6.44** (3.28)		
Section 5		-0.78 (2.16)		
10/29/2020 : Section 5			7.31** (3.44)	
11/05/2020 : Section 5			4.80 (3.68)	
11/17/2020 : Section 1			6.44** (3.28)	
11/18/2020 : Section 2			5.62 (7.30)	
11/18/2020 : Section 3			8.91*** (3.41)	
11/19/2020 : Section 5			5.74 (4.49)	
12/03/2020 : Section 5			5.17 (3.51)	
10/29/2020 : Section 4 : Study 2				-4.10 (7.88)
10/29/2020 : Section 5 : Study 1				3.23 (2.16)
10/29/2020 : Section 5 : Study 2				8.18** (4.15)
11/05/2020 : Section 5 : Study 1				-1.13 (3.31)
11/05/2020 : Section 5 : Study 2				8.03** (3.81)
11/17/2020 : Section 1 : Study 1				5.15 (3.35)
11/17/2020 : Section 1 : Study 2				4.52 (2.87)
11/18/2020 : Section 2 : Study 1				2.29 (14.07)
11/18/2020 : Section 2 : Study 2				5.73 (7.63)
11/18/2020 : Section 3 : Study 1				8.86* (4.54)
11/18/2020 : Section 3 : Study 2				5.75** (2.55)
11/19/2020 : Section 5 : Study 1				4.12 (3.85)
12/03/2020 : Section 5 : Study 1				5.39 (4.35)
12/03/2020 : Section 5 : Study 2				1.73 (2.64)
Constant	1.27 (1.42)	2.21 (2.15)	-4.24 (2.94)	-2.77 (1.93)

Observations	89	89	89	89
R2	0.02	0.07	0.08	0.15
Adjusted R2	0.005	0.01	-0.02	-0.02

Symbols: *p<0.1; **p<0.05; ***p<0.01
(B) = Blocking, Sec = Section

Table 4: Dual Channel Models (Video Frames and Audio Transcript)

Dual Channel Models				
	Treatment	Sec (B)	Sec+Date (B)	Sec+Date+Study (B)
Treatment	-2.47 (1.95)	-2.23 (2.01)	-2.21 (2.06)	-1.92 (2.25)
Section 2		-0.69 (7.08)		
Section 3		2.62 (2.70)		
Section 4		-6.57** (3.21)		
Section 5		-0.78 (2.14)		
10/29/2020 : Section 5			7.52** (3.40)	
11/05/2020 : Section 5			4.97 (3.65)	
11/17/2020 : Section 1			6.57** (3.21)	
11/18/2020 : Section 2			5.88 (7.28)	
11/18/2020 : Section 3			9.19*** (3.35)	
11/19/2020 : Section 5			5.36 (4.31)	
12/03/2020 : Section 5			5.51 (3.44)	
10/29/2020 : Section 4 : Study 2				-4.59 (7.59)
10/29/2020 : Section 5 : Study 1				3.19 (2.21)
10/29/2020 : Section 5 : Study 2				8.25** (4.18)
11/05/2020 : Section 5 : Study 1				-1.25 (3.37)
11/05/2020 : Section 5 : Study 2				8.11** (3.83)
11/17/2020 : Section 1 : Study 1				5.02 (3.36)
11/17/2020 : Section 1 : Study 2				4.50 (2.84)
11/18/2020 : Section 2 : Study 1				2.35 (14.08)
11/18/2020 : Section 2 : Study 2				5.78 (7.68)
11/18/2020 : Section 3 : Study 1				8.85* (4.56)
11/18/2020 : Section 3 : Study 2				5.91** (2.56)
11/19/2020 : Section 5 : Study 1				3.56 (3.73)
12/03/2020 : Section 5 : Study 1				5.49 (4.42)
12/03/2020 : Section 5 : Study 2				1.91 (2.51)
Constant	1.23 (1.42)	2.15 (2.13)	-4.42 (2.90)	-2.76 (1.97)
Observations	89	89	89	89
R2	0.02	0.07	0.08	0.16
Adjusted R2	0.01	0.02	-0.01	-0.01

Symbols: *p<0.1; **p<0.05; ***p<0.01
(B) = Blocking, Sec = Section

Comparing the Single Channel with Dual Channel complete models, we did not see an improvement of treatment effect nor significant difference. This is not surprising as it is common for the typically talkative students in class to be reliably talkative regardless of our treatment administration. We did not expect our treatment to affect the subject's personality characteristics. The higher weights given to subjects who speak in class has little impact on the overall results. This is because of the abundance of engaged video frames outweighs the number of times an individual talks.

Some subjects did not reply to our treatment, and we had no way of knowing if they received the treatment; we treat these subjects as non-compliers. We used the text channel to identify 27 compliers among the treatment group of 46 subjects and naturally treated all the control subjects as compliers. We computed a Complier Average Causal Effects (CACE) for the base model and received an estimate of -1.560578. We also ran a completed (blocked) model on the compliers data obtaining a non-significant CACE estimate of -1.80275.

Table 5: Single Channel Model, Dual Channel Model, Dual Compliers Model

Single Channel VS. Dual Channels VS. Dual Channel Compliers			
	Single Channel	Dual Channel	Dual Channel: Compliers
Treatment	-1.78 (2.25)	-1.92 (2.25)	-1.80 (2.35)
10/29/2020 : Section 4 : Study 2	-4.10 (7.88)	-4.59 (7.59)	0.50 (6.74)
10/29/2020 : Section 5 : Study 1	3.23 (2.16)	3.19 (2.21)	3.16 (2.55)
10/29/2020 : Section 5 : Study 2	8.18** (4.15)	8.25** (4.18)	5.44* (3.25)
11/05/2020 : Section 5 : Study 1	-1.13 (3.31)	-1.25 (3.37)	-1.78 (3.21)
11/05/2020 : Section 5 : Study 2	8.03** (3.81)	8.11** (3.83)	9.21** (4.15)
11/17/2020 : Section 1 : Study 1	5.15 (3.35)	5.02 (3.36)	5.66 (4.30)
11/17/2020 : Section 1 : Study 2	4.52 (2.87)	4.50 (2.84)	4.75 (3.17)
11/18/2020 : Section 2 : Study 1	2.29 (14.07)	2.35 (14.08)	0.61 (20.51)
11/18/2020 : Section 2 : Study 2	5.73 (7.63)	5.78 (7.68)	-0.42 (3.16)
11/18/2020 : Section 3 : Study 1	8.86* (4.54)	8.85* (4.56)	8.87* (4.58)
11/18/2020 : Section 3 : Study 2	5.75** (2.55)	5.91** (2.56)	5.89** (2.55)
11/19/2020 : Section 5 : Study 1	4.12 (3.85)	3.56 (3.73)	6.30 (4.89)
12/03/2020 : Section 5 : Study 1	5.39 (4.35)	5.49 (4.42)	5.49 (4.46)
12/03/2020 : Section 5 : Study 2	1.73 (2.64)	1.91 (2.51)	3.53* (1.91)
Constant	-2.77 (1.93)	-2.76 (1.97)	-2.82 (2.03)
Observations	89	89	70
R2	0.15	0.16	0.16
Adjusted R2	-0.02	-0.01	-0.08

Symbols: *p<0.1; **p<0.05; ***p<0.01
 (B) = Blocking, Sec = Section

Conclusion

Our work proved instrumental for tackling a real-world problem that has high practical importance and delivers important results for the research question of assessing the impact of private chat discussions on engagement of students in Berkeley classes. As of now, we have the following results:

- A prototype of a system for automated evaluation of experiments with online learning
 - Methodology, including design and automated analysis of experiments;
 - Data collection pipeline, leveraging computer vision and NLP AI for extraction of signals;
 - DNN for classification of engaged/disengaged with high accuracy;
 - Labeled dataset for training and validation of the model.
- Proof-of-concept application of the system in our class project:
 - 7 controlled experiments on w241 students across 5 sections;

- Effect of the treatment detected in most sessions with statistical significance;
- Insights from the analysis that may be used by course instructors.

Since there is no one general approach to measure engagement, our measurement system is inherently flawed for certain subjects. This effect is multiplied by the weighting technique we applied. For future implementations, we can fine tune the measurement and iteratively modify the weighting algorithm to capture all data points. Our measurement system has the potential to perform more accurately and with higher consistency compared to evaluating engagement using human judgement.

Future work

The prototype can also be extended with new metrics and integrated with Zoom or Microsoft Teams. More concrete steps for further research and development of our work are outlined below. We would like to:

- Modify the design of experiments:
 - Try using a neutrally-themed chat message as “placebo”
 - Bring more covariates to the models such as gender, age, time of class
 - Collaborate with instructors on more meaningful treatments
- Enhance the performance of engagement prediction algorithm:
 - Create more labeled samples for the training/validation dataset
 - Balance engaged with disengaged photos
 - Interpretability improvements and the addition of other signals
 - Sliding window labeling of images series for training the model
 - Sentiment analysis of speech and chat messages
 - Integration with interactive educational platforms.
- Improve reliability of data collection pipeline:
 - Name detection with privacy preserving
 - Advancing NLP methods for sentiment detection
 - 400 HTTP status response on certain valid images
 - Improving text recognition with OCR
- Conduct more experiments:
 - Generate more data for improved statistical power on aggregated results
 - Draw personalized inference with a feedback loop to the instructor

There are important applications of this work beyond education, such as virtual medical consultations or any other online meetings that involve the delivery of instructional content. Making sure that the subject understood the instructions for the schedule and expected effects of the prescribed treatment plan may help reduce readmission rates for hospitals and lead to general health improvements. We are also envisioning possibilities of applying a similar approach in the entertainment space, e.g. for personalizing feeds of online content (video or books) based on the reaction of the user; however, all these applications require serious consideration of privacy and need to be carefully assessed from the perspective of responsible AI.

References

1. P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci & U. Trautwein. "Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction" (2019). Educational Psychology Review URL: <https://link.springer.com/article/10.1007/s10648-019-09514-z>
2. O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, & C. Pari "Automated Recognition of Student Engagement using Deep Learning and Facial Expression" (2019) URL: <https://arxiv.org/abs/1808.02324>
3. J. Whitehill, Z. Serpell, Y. Lin, A. Foster and J. R. Movellan (2014) "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," in IEEE Transactions on Affective Computing, vol. 5, no. 1, pp. 86-98, 1 Jan.-March 2014, doi: 10.1109/TAFFC.2014.2316163. URL: <https://inc.ucsd.edu/mplab/wordpress/wp-content/uploads/EngagementRecognitionFinal.pdf>
4. J. Costa., M. Jung, M. Czerwinski, F. Guimbretière, T. Le and T. Choudhury. "Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-perception." (2018) CHI '18 URL: <https://www.microsoft.com/en-us/research/uploads/prod/2018/03/CHI2018-conflicts-final.pdf>
5. S. Mota and R.W. Picard, "Automated Posture Analysis for Detecting Learner's Interest Level" (2003). Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI, June, 2003. PDFTR 574. URL: <https://affect.media.mit.edu/pdfs/03.mota-picard.pdf>
6. S. Mota and R.W. Picard, "Automated Posture Analysis for Detecting Learner's Interest Level" (2003) Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI, June, 2003. PDFTR 574. URL: <https://affect.media.mit.edu/pdfs/03.mota-picard.pdf>
7. S. Mota "Automated Posture Analysis For Detecting Learner's Affective State " (2002). MIT MS Thesis, September 2002. URL: <https://affect.media.mit.edu/pdfs/02.mota.pdf>
8. A. Savelieva, T. Payne, G. Mein "AISLE: AI-Supervised Learning Environment" (2019) Berkeley MIDS w205 final course project (Spring 2019 cohort). August 2019. URL: https://drive.google.com/file/d/1Av823uPIQW8R_kdb-WSi25M_UyYualxD/view?usp=sharing
9. A. Savelieva, "Say Goodbye to Bad Meetings: how Data Science, Face API, and HoloLens can take your communication to next level" (2019) W201 Summer 19 || Group 1| Live Session Monday 6.30PM. URL: https://drive.google.com/file/d/1v1YJ5DKLH8IXccAGeuzh6COcex8_pg1T/view?usp=sharing
10. X. Zhang, Y. Sugano, M. Fritz, A. Bulling. "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation." (2017) Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). arXiv:1611.08860
11. L. Schultz & J. Sharp. "The Effect of Class Duration on Academic Performance and Attendance in an Introductory Computer Class" (2007). Information Systems Education Journal. 6. 1-7.

https://www.researchgate.net/publication/228912278_The_Effect_of_Class_Duration_on_Academic_Performance_and_Attendance_in_an_Introductory_Computer_Class

12. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.
13. Aslan, S., Mete, S.E., Okur, E., Oktay, E., Alyuz, N., Genc, U.E., Stanhill, D., Esme, A.A.: Human expert labeling process (help): Towards a reliable higher-order user state labeling process and tool to assess student engagement. Educational Technology pp. 53–59 (2017)
14. G. Bradski. The OpenCV Library (2000). Dr. Dobb's Journal of Software Tools.
15. J.R. Landis, G.G. Koch. The measurement of observer agreement for categorical data. (1977) Biometrics. 1977;33(1):159–74. <https://pubmed.ncbi.nlm.nih.gov/843571/>
16. M. McHugh. Interrater Reliability: The Kappa Statistic Biochem Med. 2012; 22(3):276-82. <https://pubmed.ncbi.nlm.nih.gov/23092060/>
17. Klizman, Robert. The Reporting of IRB Review in Journal Articles Presenting HIV Research Conducted in the Developing World. *Dev World Bioeth.* 2011 Dec; 11(3): 161–169. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3459234/>
18. Consent form - <https://forms.gle/kdBDmRmG6asoFkQt8>

Appendix: Subject Level Analysis

While the main part of the paper discusses aggregated results at the level of sections, it is equally important to understand the impacts of experiments at the level of individual subjects, as this will enable the development of personalized online educational experiences. This part of the report focuses on analyzing the data collected in this experiment from the perspective of visually detected engagement signals. They may be represented as arrays of values engaged == 1 and disengaged == 0 before and after treatment. In the following section, we are comparing collected snapshots of a subject's behaviour over 5 minutes prior to treatment to themselves for 7 minutes after the treatment (the subject at different moment is viewed as a control group and treatment group of subjects respectively).

Exploratory Data Analysis

First of all, we want to evaluate the number of data points we got per experiment subject via processing of video frames before and after administering the treatment times. [Figure A.1] shows box plots of frame counts per subject across all experiments we conducted in this study. Since we analyzed 5 minutes of video prior to treatment and 7 minutes after it, it is expected that the mean in blue plot should be lower than in the orange. Overall, as we see from this visualization, it is rare for a subject to have less than 100 frames before or after a treatment, indicating that we have a lot of data samples to run tests with high statistical power.

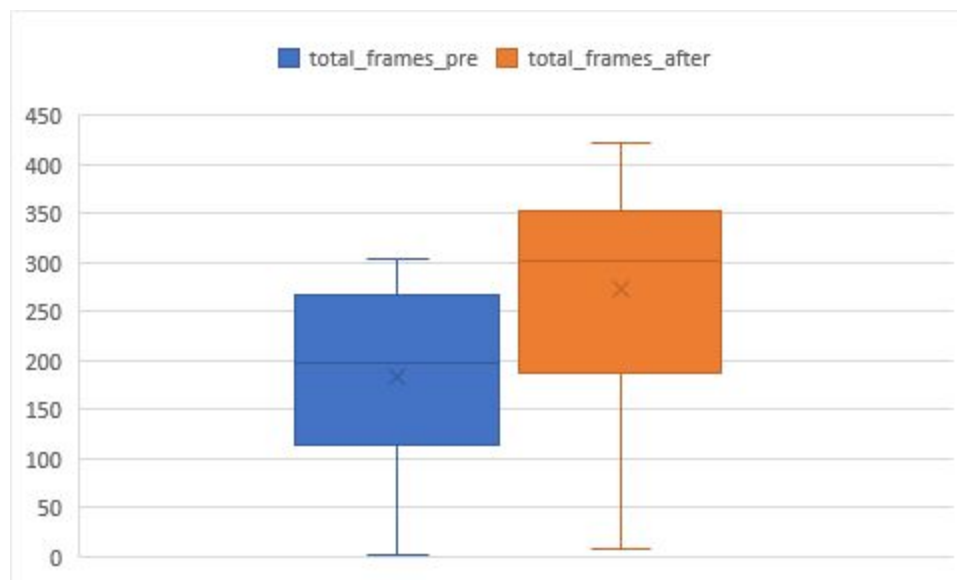


Figure A.1: Descriptive statistics for number of frames per subject in an experiment

In addition to visualizing the distribution of slides per individual, we performed data validation on random subjects to test the quality of the transformed data. As seen in [Figure A.2], we avoided

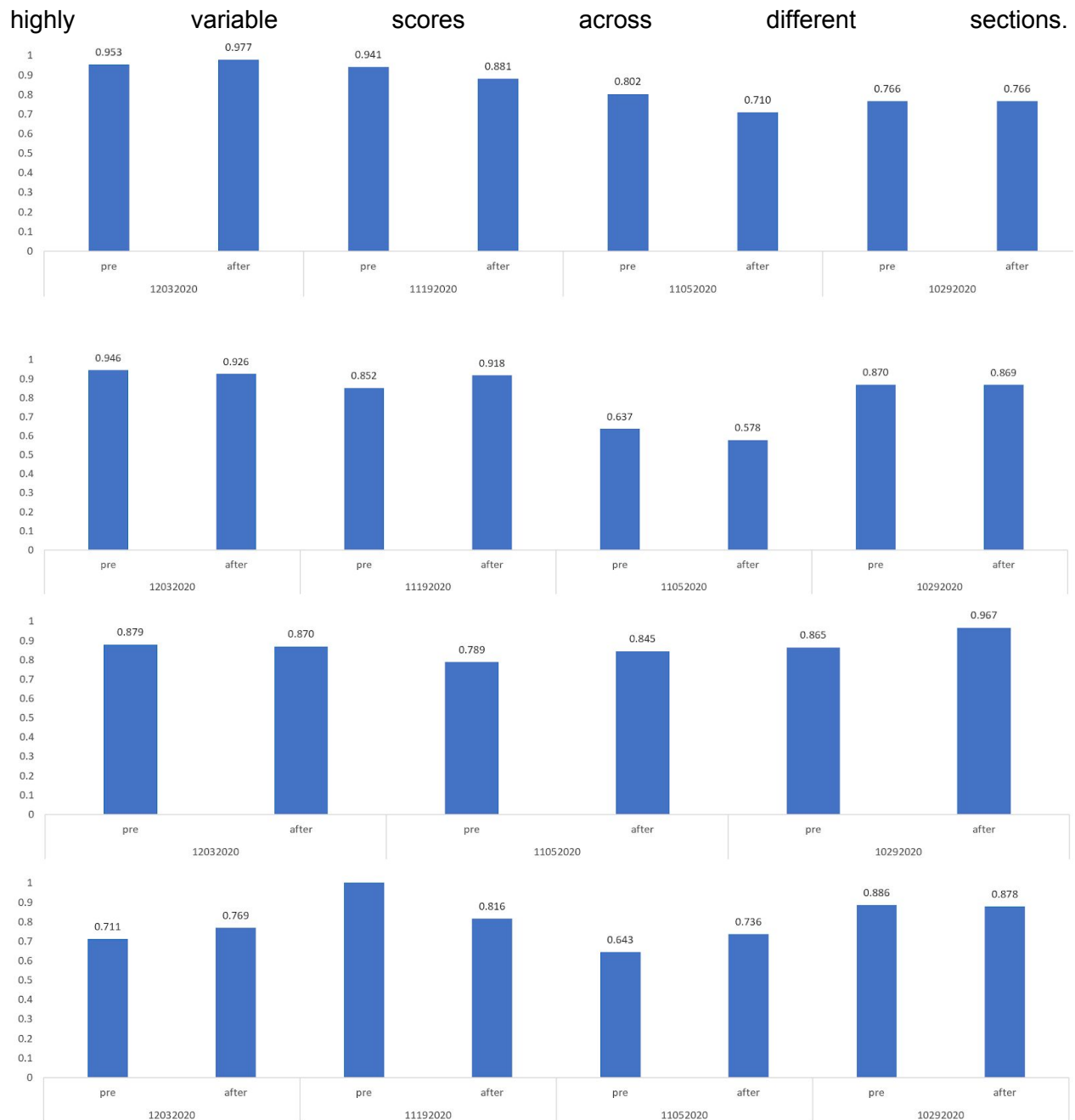


Fig. A.2. Mean of engagement signals for a random sample of four subjects in an experiment

As expected, engagement varies both across studies for the same subject and across subjects, as well as before and after the treatment. It's interesting to note that all four subjects have their lowest engagement average scores in the 11.05.2020 experiment. This suggests correlation in the group reaction to the topics being discussed and is another good sign that the data we collected is meaningful.

Statistical power analysis

Two-sample t test for means in the pre treatment and post treatment groups

We calculated power using a two-sample t-test with 28 in pre-treatment and 35 in after-treatment and a medium effect size of 0.5. The power level for a random individual with 184 pre-treatment frames and 274 post-treatment frames is 0.99. Additionally, we calculated power for the proportion of engaged versus disengaged frames using the Two-sample test for proportions. Our null hypothesis states that there is no difference in the proportion of frames that are engaged or disengaged. Our alternative hypothesis is that there is a statistically significant difference. This is a two-sided alternative. One portion of frames has a higher proportion of “engaged”. To detect a difference as small as 5% at 80% power, we found that we need a minimum of 333 slides for each group.

Statistical analysis

There are two ways to apply the main null-hypothesis at subject level. First, we can make a claim that the average treatment effect before and after treatment is the same. We can't apply the paired sample t-test, as the samples in general have different sizes. Hence, Welch t-test is used. Below is an example of output for the 3rd participant of the experiment on 11.05.2020:

```
data: x and y
t = 2.6633, df = 186.9, p-value = 0.008412
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02729056 0.18320383
sample estimates:
mean of x mean of y
0.9523810 0.8471338
```

Same test for this participant on 10.29.2020 also detects a statistically significant positive treatment effect, though with higher p-value (we can't reject null-hypothesis with 95%, only with 90%, primarily due to lower number of samples - only 37 frames in pre-treatment set):

```
data: x and y
t = 1.8516, df = 16, p-value = 0.08262
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.02556701 0.37850818
sample estimates:
mean of x mean of y
1.0000000 0.8235294
```

We ran a simulation using randomized inference in order to test the sharp null hypothesis and assess the likelihood of each unit's treatment and control potential outcome are identical. Below graphs were generated for the above examples respectively. In this setting we couldn't reject the sharp null hypothesis for the first example on either one or two-tailed tests, but got an insignificant result with $p=0.067$ on the second example. We failed to reject the null hypothesis in both scenarios, which means that we cannot reject the possibility that the proportion of engaged is equal to disengaged by chance.

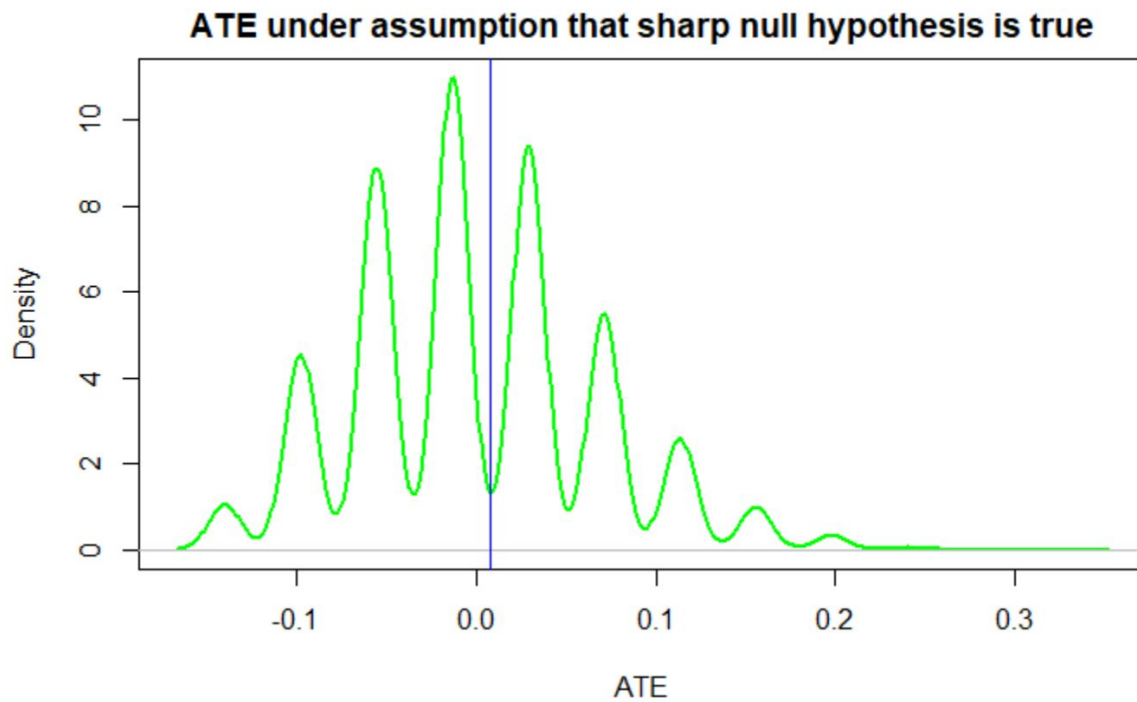


Figure A.3. RI analysis of engagement signals for one of the subjects pre & after treatment

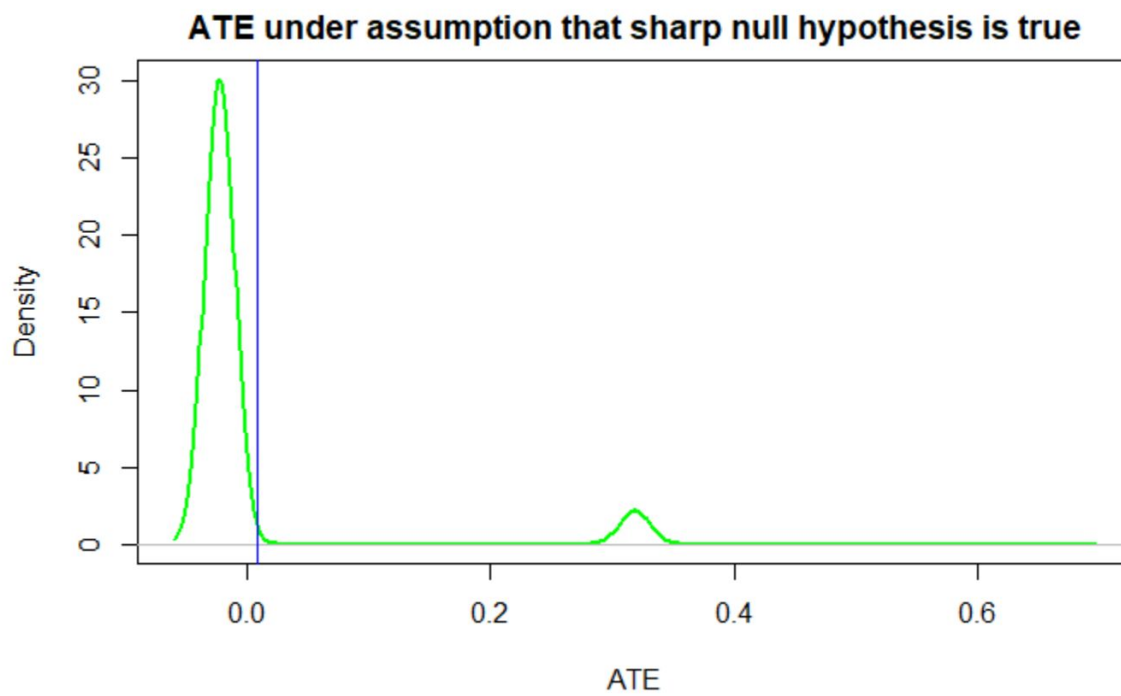


Figure A.4. RI analysis of engagement signals for one of the subjects pre & after treatment

It is worth noting that the reaction to treatment is not always positive for all subjects. For example, for the subject represented by the fourth graph in EDA we see a lack of reaction in 10.29.2020 experiment, according to the test:

```
data:  x and y
t = -0.053061, df = 628.58, p-value = 0.9577
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05232561  0.04957229
sample estimates:
mean of x mean of y
0.8684864 0.8698630
```

The same subject showed very highly statistically significant decrease in the 11.19.2020 experiment - from almost perfect 100% engagement, (s)he went down to ~82% per our model, and very high statistical significance is observed:

```
data:  x and y
t = -5.2878, df = 124, p-value = 5.402e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2528732 -0.1151268
sample estimates:
mean of x mean of y
  0.816      1.000
```

From cross-referencing these results of chat processing, we can see that the subject never responded to our treatment, that (s)he didn't notice the message and never "consumed" our treatment (but more likely, since it was our 3rd experiment in a row, is that (s)he saw it, but got irritated and chose to not reply).

Results

This study proved instrumental for tackling a real-world problem that has high practical importance and delivered statistically significant results for the research question of assessing the impact of private chat discussions on engagement of students in Berkeley classes.

The insights obtained from per-subject study are interesting for formulating research questions for future experiments, where we would like to focus on personalizing the treatment. Reinforcement learning may be used to dynamically adjust to individual characteristics of subjects.