EngageMe: A novel approach towards Multimodal Engagement Analysis in Online Learning

ANONYMOUS AUTHOR(S)*

Analyzing engagement levels in online learning is becoming increasingly significant given how classroom learning has become obsolete due to the COVID-19 pandemic. Several studies have tried to monitor students' attention levels while watching Massive Open Online Courses (MOOC) or reading an electronic document. However, most of them have relied on self-reporting or annotating the ground truth with experts' help, which creates a human bias and adds uncertainty thus becoming unfeasible in real-time due to the high annotation cost. So to overcome this ambiguity, we rely on a more definitive method for accessing ground truth attention labels using quantitative results from the clinically accepted neuropsychological tests. We propose a more generalizable approach for engagement analysis in in-the-wild settings using non-intrusive metrics such as behavioral cues and physiological signals via webcam.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: engagement, attention, mutimodal, eye-tracking, head pose, facial expressions, heart rate

ACM Reference Format:

Anonymous Author(s). 2018. EngageMe: A novel approach towards Multimodal Engagement Analysis in Online Learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Are the students engaged in a online learning environment? What are their levels of engagement? How is their engagement related to the quality of class content and learning materials? Attaining this information about the student is helpful for the instructor to enhance their engagement level. It also brings the interest of scientists from different fields like psychology and education to analyse student engagement during an online class.

In the COVID-19 pandemic situation with strict social isolation, there has been a transition in the educational system, physical classrooms have moved to online. The online education system depends on modern technologies such as webcams and microphones. This led to extensive ongoing research to detect the student's engagement in the digital learning environment using different modalities such as audio, video and physiological signals [5, 10, 14, 20]. It offers many advantages in terms of being more economical and accessible to the students. But it also brings along few challenges, such as making it difficult for the instructor to assess the student's engagement while taking the classes [49]. Thus, obtaining information about the level of engagement during an online class is a valuable asset for the instructor as well as for the students.

What is engagement? How it can be measured? Engagement is described as a multidimensional construct composed of three dimensions: emotional, behavioural, and cognitive [39]. These dimensions are dynamically interrelated factors for every student. Emotional engagement focuses on effective reactions such as happy or sad, behavioural engagement

Manuscript submitted to ACM

focuses on the action of students such as asking questions and showing interest. These engagements are externally visible whereas cognitive engagement is an internally learning process. Cognitive engagement is a psychological process involving attention and investment [21]. According to engagement theorists, engagement can be measured by two methods: external observable factors, such as body postures, eye movement, facial expressions, head pose and internal factors to the individual such as effective and cognitive function. [**?**].

Attention is a fundamental aspect of cognitive engagement. It is usually defined as a wide assortment of cognitive states, processes, and skills. Since attention is a part of cognitive engagement, in the proposed work, attention is taken as a latent variable of engagement. Levels of attention are calculated using neuroanatomic theories, factor analysis of psychometric tests, cognitive processing, and clinically based models[43]. The clinical-based model comprises five components for finding the level of attention.

- `Alternating level:` find the level of attention for mental flexibility that allows shifting the focus of attention.
- `Divided level:` find the level of attention by responding to multiple tasks at the same time.
- `Focused level:` find the level of attention by responding separately to tactile, auditory, or visual stimuli.
- `Selective level:` find the level of attention by maintaining the cognitive set during distractions or competing stimuli.
- `Sustained level:` find the level of attention by maintaining the consistent behaviour during repetitive activity.

Researchers have related the processes of selective attention to three domains that are prestigious for academic foundations: mathematics, literacy, and language. They are important for the field of education, it is both relevant for academic foundations and amenable to training [46]. Selective attention and attention switching are essential to almost all cognitive tasks [30, 37, 41]. Alternating attention is an individual's ability to change the focus of attention and the mental flexibility to shift between tasks with different cognitive requirements. This attention component is critical for students, such as when shifting between listening to lectures and writing notes [22]. Paying attention to lectures for a longer duration is an example of sustained attention in practice. [19].

Given the importance of different attention components, the proposed work aims to use different neuropsychological tests for ground truth attention labels and use effective computing techniques to measure student attention in classrooms based on non-intrusive metrics such as behavioral cues and physiological signals via webcam. Prior literature mostly utilized explicit methods of collecting feedback which induces distractions, adds workload on the audience and does not provide objective information to presenters.

Therefore for our study, we intend to cover the following goals:

- To present an EngageMe dataset, which is a multi-modal dataset of students performing online tasks and getting attention scores.
- To propose a quantitative methodology for generating reliable attention labels.
- To investigate the effectiveness of webcam-based Heart Rate (HR) calculation, facial action units, and gaze points in attention prediction.
- To explore if a non-invasive modality like an eye tracker or a combination of modalities is used to predict attention levels in students.

2 RELATED WORK

In this paper, we predict type of attention using webcam recording. Therefore, we discuss the related work in three sections: 1) Attention and Gaze, 2) Attention and facial action units, and 3) Attention and physiological signals.

2.1 Attention and Gaze

Eye movements are a widely available way and an unobtrusive method to measure cognitive load [55]. Eye movements like blink rate, pupil dilation, and the number of fixations per second are related to cognitive activity changes. And hence a lot of studies have taken it as a modality while tracking attention [1, 7, 15, 40]. In recent times various aspects of attention have been explored with eye tracking in specific tasks and accessible viewing environments. Researchers have focused on attention forecasting during interactions with a handheld mobile device using a wearable eye tracker [45], and have also explored the strategic impacts by gaze, intention, and deception [31, 32]. Researchers have also begun to use eye-tracking sensors and techniques in complex scenarios like understanding game patterns in chess, understanding the emotions induced by specific visuals, and in healthcare and market research sectors [9, 17, 23].

2.2 Attention and Facial Action Units

Though eye-tracking is the most common modality used to predict attention, facial action units (AU) are also employed by many studies [4, 28]. Researchers have tried to map differences in AU while doing a challenging task and engaged work [3]. They concluded that gaze, pitch, and lips part action units indicate engaged work, whereas pitch, gaze movements, gaze angle, and upper-lid raiser action units are indicators of challenging work. However, the other studies have fed labeled data containing facial action units into a model and then used the model to predict attention levels [48]. While most studies have explored computer vision methods to extract facial action units, some of them have also used sophisticated devices like Kinect Face Trackers to extract the features [28].

2.3 Attention and Physiological Signals

Physiological signals have been commonly used for detecting emotions, alertness, mood, task engagement,and stress. In most of the proposed methods for measuring physiological states attempt to record and analyze the signals produced by brain, heart, skin and muscles. Physiological signals like Photoplethysmography (PPG) and Electroencephalograph (EEG) are affected by the autonomic nervous system's response to emotional and psychological activity [36]. Hence, changes in these signals can detect various attentive states [34, 50, 54]. PPG can be more useful in detecting attentional states than other physiological signals [34]. In particular, these four parameters: pulse amplitude, pulse width, pulse downward slope, and mean pulse rate extracted from PPG signal, can be used to track attention levels. Heart rate can be extracted from PPG signals and can detect attention levels [54].

These studies demonstrated the potential for video-based and physiological signal-based attention detection and their generalizability, but most of these methods need contact devices to assess the student's level of attention. Using different contact devices to monitor students' attention during class may cause stress among the students which leads to poor cognitive performance of the students. To overcome this problem, an attention analysis method is proposed which takes behavioral data in-the-wild and measures the attention level of students. We are using different types of neuropsychometric tests for measuring different types of attention.

3 HYPOTHESIS DEVELOPMENT

Our general research objective is to answer two questions,

- (Question 1) Identify the relation between physiological signals and the engagement level of students during the class?

- (Question 2) Can we estimate the students' level of engagement using their behavioral data? In the proposed work, we have taken attention as a latent variable for measuring engagement.

To answer the above research objective questions, we conduct research under the following hypothesis: (Hypothesis 1) Neuropsychological tests will be able to detect attentiveness while performing complex cognitive activities which occur in a real classroom. (Hypothesis 2) Students' facial behavior and physiological features (heart rate) that are extracted using computer vision techniques are effective indicators for the measurement of attention level during any cognitive task.

For the first hypothesis, it has been considered that neuropsychological tests can measure the cognitive functions of a student during class. So, we are performing different psychometric assessment for finding the students' level of attention. Online learning activities consist of complex cognitive mechanisms which are a combination of different attention levels –sustained, divided, focused. We are using these three types of attention for measuring the level of attention. Attention abilities are usually evaluated as part of larger neuropsychological or cognitive assessments. [44].

For the second hypothesis, it has been assumed that facial expression gives enough information to measure the attention of a person [25, 51]. It can provide hints about ongoing cognitive processes and can be analyzed by considering facial action unit (AU) features. several works have been made in measuring attention from facial expressions [4, 52]. Considering the above approaches facial behavior is used as one aspect for attention assessment, it also provides a non-intrusive and continuous method for capturing facial images when the student is performing some learning tasks. The captured images are utilized to understand different aspects of students' state of mind.

4 METHODOLOGY

The training data was collected from participants while they performed an experiment explained in section 3.1. Once they completed the experiment, their webcam recording and a CSV (cognitive task responses) were uploaded from their system to the AWS S3 bucket. Both the CSV and video recording data were synchronously split into N chunks of 20 seconds each.The CSV data provided labels based on the participant's performance, while the features namely Heart Rate, Head Pose, Facial Expressions, and Eye Tracking are extracted from each clip.

4.1 Data Collection

60 participants were recruited to participate in our study (42 males, 18 females) aged between 19 and 21 years (M = 19.97, SD = 0.56). All participants were undergraduate students at IIIT Delhi. Fig 1 summarizes the study's online data collection framework[1] . The data collection experiment link was sent to the participants online, and they completed it from their homes. No compensation was provided to any participant, and informed consent was obtained before the experiment. No specialized hardware or software installation was required. Participants just needed to allow access to their webcams. Consequently, the experiment's experience was similar to their normal study environment.

The experiment was created using jsPsych [6]. It took approximately 20 minutes for each participant. As shown in fig 1, an eye-tracker calibration was followed by 3 cognitive tasks in the experiment. Each of the tasks corresponds to a particular type of attention.

4.1.1 Training Data.

[1]https://www.specialeduneeds.com/

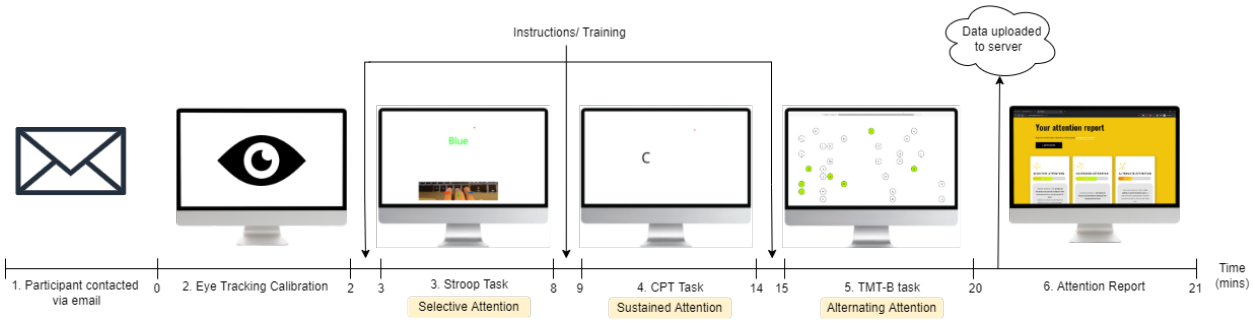


Fig. 1. Experiment timeline

- Eye tracking: A webcam-based eye tracker (webgazer.js) [33] was used in the experiment. The eye tracker needs to be calibrated first before starting the cognitive tasks. Eye-tracker validation was also added which made sure that the webgazer's accuracy is greater than 60%.
- Stroop Test: As a measure of selective attention, the Stroop color-word task is one of the most widely used cognitive tasks [53]. J.R. Stroop et al. [47] reported how it takes longer for people to verbally articulate the font colour of a word that represents an incongruent colour (the word reads "red," but it's written in "green" colour.We used the computerized version of the Stroop test, which requires the participant to use the arrow keys to report the font colour of the text displayed, regardless of its meaning. Each stimulus remained on the screen for 4000 msec or until a response was collected. [35]. The post-trial gap after each stimulus was also kept variable between 100-500 milliseconds. Training trials were also provided, with the same distribution as the test experimental blocks. A feedback signal on the error and correct responses were provided to facilitate the task learning. Feedback was absent in the actual experiment [29], and the texts were displayed anywhere on the screen to match its coordinates with the webgazer response in post-processing. The data from the training were not analyzed.
- Continuous Performance Test (CPT): The CPT test is a widely used measure for sustained attention [11] We used the N-back CPT test (N=2) as a measure for sustained attention and working memory. The n-back task requires participants to respond to a series of stimuli by judging whether each stimulus is the same as the one presented n-items previously.[16]. In our experiment, the participants had a sequence of letters appearing on the screen, and they had to press the "M" key whenever the current alphabet was the same as the one presented 2 items back, for example, A, B, A. The stimulus duration was 1500 ms, and there was a 500 ms gap post-trial. Training trials were provided and actual test had the coordinates of the letters changing on the screen.
- Trail Making Test (TMT): The Trail Making Test (TMT) is a frequently used neuropsychological test for assessing cognitive performance and alternating attention [44]. Conventionally this task is done on pen and paper, but we have implemented a computerized version. There are two types of TMT tests, i.e., A and B. TMT-A involves number sequencing (1 to 15), whereas TMT-B includes set-shifting: it requires the subject to alternate between numerical and alphabetic sequences (1–A–2–B–3...)[38]. We used only the TMT-B task for analyzing the participant's alternating attention for our study. Three trials for the TMT-B task were conducted.

‘

4.2 Labels generation from CSVs

4.2.1 Data preprocessing. :

Before performing any task-specific analysis, all the distributions received (RTs, accuracy, total time) for the 3 tests were checked for normality using the Kolmogorov-Smirnov test [35]. Z-score analysis was used to remove outliers having an absolute z-score > 3 in their respective column (<5% of the total trials were excluded).

Stroop: Paired t-test showed significant condition effect ($t(2,59)=1.865$, $P<0.05$) in the congruent and incongruent stimuli. The mean of the median RT for all participants for the congruent condition was 825 ms, while the same for the incongruent condition was 895 ms.

CPT: In CPT, the number of correct hits is considered an indication of attention. In contrast, the two types of errors, i.e. omission (missing a target) and commission (responses to stimuli other than the target), are considered measures for inattention and impulsivity, respectively [13]. Due to impulsivity, a negative correlation is expected in the number of false alarms and the median RTs. Pearson's correlation of (0.357, $P<0.05$) was found between the two mentioned parameters indicating commission errors. Moreover, the average number of commission errors in the best 50% of the participants (ranked on the basis of RT) was found 7.3 as compared to 4.2 in the rest.

TMT: As the trial progresses in the trail-making test, it gets easier to respond because more bubbles turn green and fewer competing stimuli are left. This leads to a decreased dwell time as the test progresses [38]. We examined the first and second half of the 3 TMT trials to verify the same in our data. Paired T-test on the two distributions gave ($t(2,58) = 4.50$, $p<0.01$), ($t(2,58) = 3.78$, $p<0.01$) and ($t(2,58) = 6.93$, $p<0.01$) respectively for the three trials. Mean Response time (RT) (of all three trials) for the first half (2899 ms) was found significantly greater than that of the second half (2141 ms).

4.2.2 Performance metrics. :

Different scoring metrics were used to rate participants' performance in the tests. The dataset (both CSVs and videos) was split into windows of 20 seconds for Stroop and CPT with a 50% overlap. Hence, from a data of 5 minutes of Stroop, we generated approximately 30 splits. For TMT, each trial was considered as a data point. Hence, three trials generated 3 data points. Approximately 500-600 splits were generated from each participant's data.

Stroop: Stroop test had two types of stimuli, i.e. congruent and incongruent. The scoring metric used was the inverse of the well-known IE score (IE = mean RT for a particular condition divided by the proportion of correct responses for that condition) that provides some correction for potential speed-accuracy trade-offs present in the data [18, 29, 42]. A lower IE score implies better performance, and as we've used the inverse IE, a higher score indicated the same in our case. We've called the inverse of IE score as the efficiency score (accuracy/ median RT). So, the efficiency score for each participant in both conditions was calculated. Both the calculated scores were scaled in 0 to 1 and then added together to give a Stroop score in the range 0-2, where 2 is the best and 0 being the worst. In cases of missing data, the participant's mean performance was considered.

CPT: Efficiency score was found for the match cases (trials with the pattern on screen) in CPT as well. But only the accuracy was considered for the rest of the non-matching cases where no response was recorded. This efficiency score and accuracy were again scaled and added to give a final CPT score in the range 0-2. The match cases were missing in some of the 20-second splits; the overall mean efficiency was considered there.

TMT: The total time taken in TMT trials was considered as the performance metric [38]. The total time taken in each trial was scaled to give a score in the range 0-1, where a lesser score implied lesser time and hence, a better performance.

4.2.3 *Labels.* :

Hierarchical labels were generated where each data point had one label for the level of attention while the other label was for the type of attention. Labels were decided based on performance scores to facilitate a multi-labeled classification model of attention. The labels based on the performance were given follows:

- `-1` : Inattention.
- `0` : Low attention.
- `1` : Moderate attention.
- `2` : High attention.

The data points which were more than 3 SD away from the overall dataset's mean were labeled -1 (inattentive). The rest of the data was divided into quartiles, and the data points in the first quartile (worst 25%) were labeled as 0 (low attention), the second and third quartile as 1 (moderately attentive), and the fourth quartile (best performing 25%) as 2 (highly attentive). Hence, three levels (0,1,2) were possible for each type of attention (sustained, selective and alternating) and a label -1 was there for inattention. Fig 2 shows the distribution of labels for Stroop test.

Label	Count
-1	26
0	230
1	307
2	231

Fig. 2. Distribution of the labels for Selective attention

4.3 Feature Extraction

We constructed the following pipelines to extract, validate and preprocess features from clips of different neuropsychological tests:

1. Eye-tracking We extracted eye-gaze features from WebGazer.js [33] for each clip. The WebGazer generates eye-gaze locations around 30 frames per second (fps). But, we found these to be in the range of 16 to 29 fps for our dataset. Moreover, the generated eye-gaze locations are unsuitable to estimate eye-tracking features namely saccade(s) as its minimum fps requirement is 60. To overcome this challenge and maintain consistency across each clip, we interpolated the eye-gaze locations using a Kalman filter. A heatmap of eye-gaze locations after processing with Kalman filter is shown in the figure 3. We validated the eye-gaze locations by estimating the Mean Euclidean Distance (mean euc) from the target centre to the gaze coordinates. Table 1 shows the Mean EUC for all clips of each of the neuropsychological tests. These eye-gaze locations were then processed to generate eye-tracking features. Table 2 shows the processed eye-tracking features.

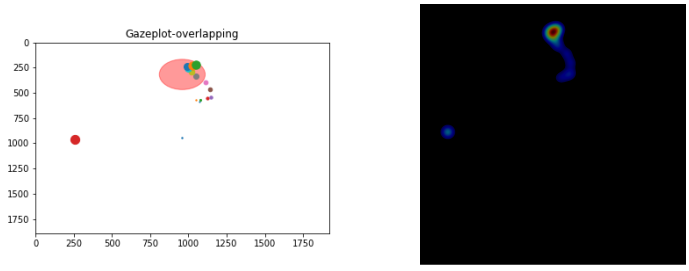
2. Head Pose and Facial Expressions We extracted head pose and facial features from OpenFace[2]. OpenFace generated 712 features for each frame in a clip. The features are related to Gaze (gaze angles, location of 2D/3D eye region landmarks), Pose (head position and head rotations), Facial Action Units (AUs, their presence and intensity), Landmarks locations in 2D, Landmarks locations in 3D and rigid and non-rigid shape parameters. We discarded the data for users where the output for success (successful tracking of the face in the frame) was 0 for more than 500 frames

Attention Label	Mean EUC
-1 (inattention)	74.85
0 (low attention)	50.08
1 (moderate attention)	43.27
2 (high attention)	34.98

Table 1. Mean EUC for all clips of each of the neuropsychological tests

Eye movement parameters	Extracted features
Fixation duration (ms)	Mean, standard deviation
Saccade	Mean, standard deviation of saccade duration (ms) and saccade amplitude
Event statistics	fixation frequency, mean fixation duration, total fixation dispersion, time to first fixation, saccade frequency, average saccade duration, average saccade amplitude, and average saccade latency

Table 2. Eye Tracking Features



(a) Gazeplot overlapping with stimulus (b) Gaze heatmap

Fig. 3. Processed gaze : (a) Shows the gazeplot obtained from a user overlapping with the area of interest. (b) denotes the heatmap of the gazeplot where red colour represents highest concentration of gaze points.

for multiple clips. This represented the cases where the subject's complete face wasn't captured. We used only Pose and AUs related features from all the generated features as we had a completely separate pipeline for Gaze related features. As Pose or AUs change only over a period of time, we calculated their mean representation per second for each clip. We preprocessed these outputs and computed features for each clip.

3. Heart Rate (HR) We extracted remote photoplethysmography (rPPG) signals from MTTS-CAN [26] for each clip. MTTS-CAN generates rPPG signals as well as respiration signals but only the rPPG signals were used. The rPPG signals were processed to estimate HR using NeuroKit [27]. To estimate the HR, we detected the systolic peaks using the Elgendi method [8] from the rPPG signals. The Kubios algorithm [24] was then used to identify and correct erroneous peak placements. These correct systolic peaks were then used to calculate the RR intervals. The RR interval and HR are hyperbolically related (HR x RR interval = 60000) [12]. Hence, HR was calculated as 60000 divided by the mean RR interval for each clip. We even validated the HR by calculating it for three subjects using fingertip reflectance

photoplethysmography (PPG) signals obtained from Shimmer3 GSR+. Table 3 shows the RMSE between the HR obtained from rPPG signals of MTTS-CAN and the PPG signals of Shimmer3 GSR+ for all the clips of each neuropsychological test.

Neuropsychological Test	RMSE
Stroop	11.45
CPT	10.72
TMT	10.20

Table 3. Root Mean Squared Error for HR obtained from rPPG signals of MTTS-CAN and the PPG signals of Shimmer3 GSR+

5 CONCLUSION AND FUTURE WORK

This paper developed a protocol to generate quantitatively backed labels for attention using neurophysiological tests. We designed a pipeline to extract, validate and preprocess the data collected in natural settings using only a webcam. We first plan to explore multi-modal deep learning architectures preserving the hierarchal dimension of attention labels generated in future work. Second, we plan to test the model on data from an already developed protocol that consists of real-world parallels for attention categories.

REFERENCES

[1] Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying Attention Types with Thermal Imaging and Eye Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 69 (sep 2019), 27 pages. https://doi.org/10.1145/3351227

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[3] Ebrahim Babaei, Namrata Srivastava, Joshua Newn, Qiushi Zhou, Tilman Dingler, and Eduardo Velloso. 2020. Faces of focus: A study on the facial cues of attentional states. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[4] Nigel Bosch and Sidney K. D'Mello. 2021. Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing* 12, 4 (2021), 974–988. https://doi.org/10.1109/TAFFC.2019.2908837

[5] Nigel Bosch, Sidney K. D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting Student Emotions in Computer-Enabled Classrooms. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 4125–4129.

[6] Joshua R De Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47, 1 (2015), 1–12.

[7] Sidney D'Mello, Kristopher Kopp, Robert Earl Bixler, and Nigel Bosch. 2016. Attending to Attention: Detecting and Combating Mind Wandering during Computerized Reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1661–1669. https://doi.org/10.1145/2851581.2892329

[8] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. 2013. Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions. *PLoS ONE* 8, 10 (Oct. 2013), e76585. https://doi.org/10.1371/journal.pone.0076585

[9] Maite Frutos-Pascual and Begonya Garcia-Zapirain. 2015. Assessing visual attention using eye tracking sensors in intelligent cognitive therapies based on serious games. *Sensors* 15, 5 (2015), 11092–11117.

[10] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O. Zander. 2014. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience* 8 (2014). https://doi.org/10.3389/fnins.2014.00385

[11] Farnaz Ghassemi, Mohammad Moradi, Mehdi Tehrani-Doost, and Vahid Abootalebi. 2009. Classification of sustained attention level based on morphological features of EEG's independent components. *2009 ICME International Conference on Complex Medical Engineering, CME 2009*, 1 – 6. https://doi.org/10.1109/ICCME.2009.4906628

[12] Jeffrey J. Goldberger, Nils P. Johnson, Haris Subacius, Jason Ng, and Philip Greenland. 2014. Comparison of the physiologic and prognostic implications of the heart rate versus the RR interval. *Heart Rhythm* 11, 11 (Nov. 2014), 1925–1933. https://doi.org/10.1016/j.hrthm.2014.07.037

[13] Jeffrey M Halperin, Lorraine Wolf, Edward R Greenblatt, and Gerald Young. 1991. Subtype analysis of commission errors on the continuous performance test in children. *Developmental neuropsychology* 7, 2 (1991), 207–217.

[14] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5114–5119. https://doi.org/10.1145/3025453.3025669

[15] Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Masha R Jones, Benjamin Katz, Martin Buschkuehl, Susanne M Jaeggi, and Priti Shah. 2020. Exploring N-Back cognitive training for children with ADHD. *Journal of attention disorders* 24, 5 (2020), 704–719.

[17] Jennifer L Kellough, Christopher G Beevers, Alissa J Ellis, and Tony T Wells. 2008. Time course of selective attention in clinically depressed young adults: An eye tracking study. *Behaviour research and therapy* 46, 11 (2008), 1238–1243.

[18] Steffan Kennett, Martin Eimer, Charles Spence, and Jon Driver. 2001. Tactile-Visual Links in Exogenous Spatial Attention under Different Postures: Convergent Evidence from Psychophysics and ERPs. *Journal of Cognitive Neuroscience* 13, 4 (2001), 462–478. https://doi.org/10.1162/08989290152001899

[19] Li-Wei Ko, Oleksii Komarov, W. David Hairston, Tzyy-Ping Jung, and Chin-Teng Lin. 2017. Sustained Attention in Real Classroom Settings: An EEG Study. *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00388

[20] Tanja Krumpe, Christian Scharinger, Wolfgang Rosenstiel, Peter Gerjets, and Martin Spüler. 2018. Unity and diversity in working memory load: Evidence for the separability of the executive functions updating and inhibition using machine learning. *Biological Psychology* 139 (2018), 163–172. https://doi.org/10.1016/j.biopsycho.2018.09.008

[21] Sergej Lackmann, Pierre-Majorique Léger, Patrick Charland, Caroline Aubé, and Jean Talbot. 2021. The influence of video format on engagement and performance in online learning. *Brain Sciences* 11, 2 (2021), 128. https://doi.org/10.3390/brainsci11020128

[22] Yi-Jung Lai and Kang-Ming Chang. 2020. Improvement of Attention in Elementary School Students through Fixation Focus Training Activity. *International Journal of Environmental Research and Public Health* 17, 13 (2020), 4780. https://doi.org/10.3390/ijerph17134780

[23] Jian Li, Li Su, Bo Wu, Junbiao Pang, Chunfeng Wang, Zhe Wu, and Qingming Huang. 2016. Webpage saliency prediction with multi-features fusion. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 674–678.

[24] Jukka A. Lipponen and Mika P. Tarvainen. 2019. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. *Journal of Medical Engineering & Technology* 43, 3 (April 2019), 173–181. https://doi.org/10.1080/03091902.2019.1640306

[25] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. 2011. The Computer Expression Recognition Toolbox (CERT). *Face and Gesture 2011* (2011). https://doi.org/10.1109/fg.2011.5771414

[26] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. *arXiv preprint arXiv:2006.03790* (2020).

[27] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53, 4 (feb 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[28] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. 2017. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing* 8, 1 (2017), 15–28. https://doi.org/10.1109/TAFFC.2016.2515084

[29] Stefanie Mühlberg, Giovanni Oriolo, and Salvador Soto-Faraco. 2014. Cross-modal decoupling in temporal attention. *European Journal of Neuroscience* 39, 12 (2014), 2089–2097.

[30] Peter Mundy, Jessica Block, Christine Delgado, Yuly Pomares, Amy Vaughan Van Hecke, and Meaghan Venezia Parlade. 2007. Individual Differences and the Development of Joint Attention in Infancy. *Child Development* 78, 3 (2007), 938–954. https://doi.org/10.1111/j.1467-8624.2007.01042.x

[31] Joshua Newn, Fraser Allison, Eduardo Velloso, and Frank Vetere. 2018. Looks Can Be Deceiving: Using Gaze Visualisation to Predict and Mislead Opponents in Strategic Gameplay. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173835

[32] Joshua Newn, Eduardo Velloso, Fraser Allison, Yomna Abdelrahman, and Frank Vetere. 2017. Evaluating Real-Time Gaze Representations to Infer Intentions in Competitive Turn-Based Strategy Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Amsterdam, The Netherlands) *(CHI PLAY '17)*. Association for Computing Machinery, New York, NY, USA, 541–552. https://doi.org/10.1145/3116595.3116624

[33] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, 3839–3845.

[34] María Dolores Coca Peláez, María Teresa Lozano Albalate, Alberto Hernando Sanz, Montserrat Aiger Vallés, and Eduardo Gil. 2018. Photoplethysmographic waveform versus heart rate variability to identify low-stress states: Attention test. *IEEE journal of biomedical and health informatics* 23, 5 (2018), 1940–1951.

[35] Raveendranadh Pilli, MUR Naidu, Usha Rani Pingali, JC Shobha, and A Praveen Reddy. 2013. A computerized Stroop test for the evaluation of psychotropic drugs in healthy participants. *Indian journal of psychological medicine* 35, 2 (2013), 180.

[36] Stephen W Porges. 2003. The polyvagal theory: Phylogenetic contributions to social behavior. *Physiology & behavior* 79, 3 (2003), 503–513.

[37] Michael I. Posner and Mary K. Rothbart. 2007. Research on Attention Networks as a Model for the Integration of Psychological Science. *Annual Review of Psychology* 58, 1 (2007), 1–23. https://doi.org/10.1146/annurev.psych.58.110405.085516

[38] Alexander Prange, Michael Barz, Anika Heimann-Steinert, and Daniel Sonntag. 2021. *Explainable Automatic Evaluation of the Trail Making Test for Dementia Screening*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445046

[39] K. Ann Renninger and Jessica E. Bachrach. 2015. Studying triggers for interest and engagement using observational methods. *Educational Psychologist* 50, 1 (2015), 58–69. https://doi.org/10.1080/00461520.2014.999920

[40] Tarmo Robal, Yue Zhao, Christoph Lofi, and Claudia Hauff. 2018. Webcam-based Attention Tracking in Online Learning: A Feasibility Study. 189–197. https://doi.org/10.1145/3172944.3172987

[41] M. Rosario Rueda, Michael I. Posner, and Mary K. Rothbart. 2005. The Development of Executive Attention: Contributions to the Emergence of Self-Regulation. *Developmental Neuropsychology* 28, 2 (2005), 573–594. https://doi.org/10.1207/s15326942dn2802_2

[42] David I. Shore, Morgan E. Barnes, and Charles Spence. 2006. Temporal aspects of the visuotactile congruency effect. *Neuroscience Letters* 392, 1-2 (2006), 96–100. https://doi.org/10.1016/j.neulet.2005.09.001

[43] Mckay Moore Sohlberg and Catherine A. Mateer. 1987. Effectiveness of an attention-training program. *Journal of Clinical and Experimental Neuropsychology* 9, 2 (1987), 117–130. https://doi.org/10.1080/01688638708405352

[44] McKAY MOORE SOHLBERG and CATHERINE A. MATEER. 2006. Improving attention and managing attentional problems. *Annals of the New York Academy of Sciences* 931, 1 (2006), 359–375. https://doi.org/10.1111/j.1749-6632.2001.tb05790.x

[45] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting User Attention during Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. https://doi.org/10.1145/3229434.3229439

[46] Courtney Stevens and Daphne Bavelier. 2012. The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience* 2 (2012). https://doi.org/10.1016/j.dcn.2011.11.001

[47] J. R. Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18, 6 (1935), 643–662. https://doi.org/10.1037/h0054651

[48] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2021. Multimodal Engagement Analysis from Facial Videos in the Classroom. *IEEE Transactions on Affective Computing* (2021), 1–1. https://doi.org/10.1109/TAFFC.2021.3127692

[49] B. Jill Venton and Rebecca R. Pompano. 2021. Strategies for enhancing remote student engagement through active learning. *Analytical and Bioanalytical Chemistry* 413, 6 (2021), 1507–1512. https://doi.org/10.1007/s00216-021-03159-0

[50] S. Ward, M. Brickley, J. Sharry, G. McDarby, and C. Heneghan. 2004. Assessment of heart rate and electrodermal activity during sustained attention to response tests. In *Computers in Cardiology, 2004.* 473–476. https://doi.org/10.1109/CIC.2004.1442977

[51] Jacob Whitehill, Marian Bartlett, and Javier Movellan. 2008. Automatic facial expression recognition for intelligent tutoring systems. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2008). https://doi.org/10.1109/cvprw.2008.4563182

[52] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98. https://doi.org/10.1109/taffc.2014.2316163

[53] Barlow C. Wright. 2016. What Stroop tasks can tell us about selective attention from childhood to adulthood. *British Journal of Psychology* 108, 3 (2016), 583–607. https://doi.org/10.1111/bjop.12230

[54] Xiang Xiao and Jingtao Wang. 2017. Understanding and Detecting Divided Attention in Mobile MOOC Learning. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2411–2415. https://doi.org/10.1145/3025453.3025552

[55] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2018. Studying Eye Movements as a Basis for Measuring Cognitive Load. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188628