n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild

NAN GAO, RMIT University, Australia
WEI SHAO, RMIT University, Australia
MOHAMMAD SAIEDUR RAHAMAN, RMIT University, Australia
FLORA D. SALIM, RMIT University, Australia

The study of student engagement has attracted growing interests to address problems such as low academic performance, disaffection, and high dropout rates. Existing approaches to measuring student engagement typically rely on survey-based instruments. While effective, those approaches are time-consuming and labour-intensive. Meanwhile, both the response rate and quality of the survey are usually poor. As an alternative, in this paper, we investigate whether we can infer and predict engagement at multiple dimensions, just using sensors. We hypothesize that multidimensional student engagement level can be translated into physiological responses and activity changes during the class, and also be affected by the environmental changes. Therefore, we aim to explore the following questions: *Can we measure the multiple dimensions of high school student's learning engagement including emotional, behavioural and cognitive engagement with sensing data in the wild? Can we derive the activity, physiological, and environmental factors contributing to the different dimensions of student learning engagement? If yes, which sensors are the most useful in differentiating each dimension of the engagement?* Then, we conduct an in-situ study in a high school from 23 students and 6 teachers in 144 classes over 11 courses for 4 weeks. We present the *n-Gage*, a student engagement sensing system using a combination of sensors from wearables and environments to automatically detect student in-class multidimensional learning engagement. Extensive experiment results show that *n-Gage* can accurately predict multidimensional student engagement in real-world scenarios with an average mean absolute error (MAE) of 0.788 and root mean square error (RMSE) of 0.975 using all the sensors. We also show a set of interesting findings of how different factors (e.g., combinations of sensors, school subjects, CO_2 level) affect each dimension of the student learning engagement.

CCS Concepts: • **Human-centered Computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Education.*

Additional Key Words and Phrases: Engagement Prediction, Students, Wearable, Electrodermal Activity

ACM Reference Format:
Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D. Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 79 (September 2020), 26 pages. https://doi.org/10.1145/3411813

1 INTRODUCTION

In education, *student engagement* refers to the degree of attention, interest, curiosity, and involvement in the learning environment [43]. The study of student engagement has attracted growing interests as a way to address

Authors' addresses: Nan Gao, nan.gao@rmit.edu.au, RMIT University, Melbourne, Australia, 3000; Wei Shao, wei.shao@rmit.edu.au, RMIT University, Melbourne, Australia, 3000; Mohammad Saiedur Rahaman, saiedur.rahaman@rmit.edu.au, RMIT University, Melbourne, Australia, 3000; Flora D. Salim, flora.salim@rmit.edu.au, RMIT University, Melbourne, Australia, 3000.

2474-9567/2020/9-ART79 $15.00
https://doi.org/10.1145/3411813

the problems of low academic achievement, high levels of student boredom, disaffection, and high dropout rates in urban areas [34, 35]. Previous research showed that student engagement declines as students progress from elementary to middle school, reaching its lowest levels in high school [22, 58, 59]. Marks et al. [58] estimated that as many as 40-60% of high school students are disengaged (e.g., uninvolved, no interests, and not attentive). The consequences of disengagement for high school students are severe. They are less likely to graduate from high school and face limited employment prospects, increasing risks for poverty, poorer health, and involvement in the criminal justice system [22]. Given the negative impact of disengagement, more and more researchers, educators, and policymakers are interested in obtaining data on student engagement and disengagement for needs assessment, diagnosis, and preventive measures [59].

Generally, student engagement is defined as a meta-construct that includes three dimensions [34, 35]: (1) *behavioural engagement* focuses on participation and involvement in academic, social, and co-curricular activities. Some researchers define behavioural engagement with regards to positive conduct, e.g., following the rules, and the absence of disruptive behaviour such as skipping school [32, 33, 35]; (2) *emotional engagement* focuses on the extent of positive and negative reactions to teachers, classmates, academics, and school, which includes a sense of belonging or connectedness to the school [31, 35]; (3) *cognitive engagement* draws on the idea of investment in learning. It incorporates thoughtfulness and willingness to put effort to comprehend complex ideas and master difficult skills [21, 34, 35]. One of the widely used method for measuring student engagement is self-report survey, e.g., *Motivated Strategies for Learning Questionnaire* (MSLQ) [74], *School Engagement Measure* (SEM) [64], and *Engagement vs. Disaffection with Learning* (EvsD) [87]. Though generally reliable, the survey is time-consuming and may become a burden for participants if they need to complete it for each class.

Therefore, we want to investigate whether we can infer and predict multidimensional student engagement just using sensors. In particular, we conduct the research around the following hypothesis: the multidimensional student engagement level can be translated into physiological responses and activity movements during the class, and can also be affected by environmental changes. In previous studies, various physiological data, (e.g., electrodermal activity (EDA), heart rate variability (HRV), accelerometer (ACC), skin temperature (ST)) and environmental data have been explored to assess the emotional arousal and engagement level in different scenarios. For instance, EDA is usually considered as a good indicator of psychological or physiological arousal (e.g., emotional and cognitive states [12, 23]) , which has been increasingly explored in affective computing, such as the detection of emotion [7, 16], depression [79] , and engagement [27, 47, 55]. Recently, Pflanzer et al. [73] stated that EDA monitoring should be combined with the heart rate because they are both autonomically dependent variables. Heart rate has been used for student engagement prediction [63] and the correlation of heart rate and cognitive/emotional engagement has been found in [37]. As the most commonly used sensor in IoT devices, accelerometer is proven to be powerful for quantifying human behavioural patterns [39, 93]. It has been used for demonstrating how synchronized movement of people can enhance group affiliation [91] and sensing engagement using interpersonal movement synchrony [95].

In this paper, our research questions are as follows: 1. *Can we measure the multiple dimensions of high school student's learning engagement including emotional, behavioural and cognitive engagement with sensing data in the wild?* 2. *Can we derive the activity, physiological, and environmental factors contributing to the different dimensions of student learning engagement? If yes, which sensors are the most useful in differentiating each dimension of the engagement?* To answer the above questions and enable automated engagement detection, we present a new classroom sensing system *n-Gage* to assess the behavioural, emotional and cognitive engagement levels of high school students. The system utilizes sensing data from two sources: (1) wearable devices capturing physiological and physical signals (e.g., EDA, HRV, ACC, ST); (2) indoor weather stations capturing environmental changes (e.g., temperature, CO_2, sound). The study has been approved by the Human Research Ethics Committee at our University and the high school where it is conducted, and all the procedures follow the ethical codes. The main contributions of this paper include:

- We collect a dataset of 23 high school students and 6 teachers in 144 classes over 11 courses for 4 weeks. Weather stations are installed in 3 classrooms, and student participants are asked to wear the E4 wristbands and complete online survey 3 times a day to report their behavioural, emotional and cognitive engagement level during the classes. To the best of our knowledge, this is the most diverse and largest dataset collected in the wild to measure student engagement using sensors.
- We build *n-Gage*, a classroom sensing system to automatically measure the multidimensional engagement (behavioural, emotional and cognitive engagement) of high school students during the classes. In particular, we combine physiological signals, physical activities, and indoor environmental data to estimate the changes in student engagement levels. To the best of our knowledge, this is the first system to detect student engagement from multiple sensors in the wild.
- We extract new features to represent the physiological and physical synchrony between students which proved to be useful for the student engagement prediction. For the first time, we extract features from skin temperature and indoor environment for effective engagement estimation.
- We conduct comprehensive experiments to predict multidimensional student engagement scores with *LightGBM* regressors. The experiment results show that *n-Gage* reaches a high accuracy (0.563 MAE and 0.715 RMSE score) for student engagement prediction. We also derive different factors and explore the most useful sensors in differentiating each dimension of the learning engagement.
- We show a set of interesting insights on how different factors affect student engagement. For example, the CO_2 level in the classroom has a negative impact on students' cognitive engagement, which highlights the need to ventilate the classroom timely to improve student engagement.

The remainder of the paper is as follows. Section 2 introduces related works of traditional methods for measuring engagement, and the recent progress of engagement prediction with sensing technology. Section 3 describes the data collection procedures, including participant recruitment, equipments for data collection, and the self-report instrument. Then we introduce data pre-processing techniques in Section 4. In Section 5, we extract various features from physiological signals and environmental changes. Section 6 introduces the prediction pipeline and Section 7 shows experiment results and in-depth discussion about engagement prediction. Section 8 lists the implications and limitations of our work. Finally, we summarize this research in Section 9 and indicate the potential direction of future work.

2 RELATED WORK

2.1 Traditional Methods for Measuring Engagement

In the education area, there are various methods to study student engagement. (1) *Student Self-report* is the most common method to assess student engagement as it is easy to execute in classroom settings. Students are provided with items reflecting different dimensions of engagement and then select the response that best describes them [35]. However, the self-report survey is labour and time-consuming, and students may not be willing to answer too many questions honestly at a time, leading to low-quality responses [5]. (2) *Experience Sampling* [85] allows researchers to collect responses at the moment, which reduces the problems of recall-failure and social-desirability bias happened in the self-report surveys. However, it requires a huge time investment from students, and the quality of responses largely relies on the students' willingness and ability to answer [35]. (3) *Teacher Ratings of Students* [35] can be useful for young students with difficulty in completing self-report surveys. Behaviour can be observed directly from teachers, but emotion engagement is difficult to be observed as students may learn to mask their emotions [35, 86]. (4) *Interviews* can provide a detailed description of the student's performance during the learning process. However, the quality of responses depends on the expert knowledge from the interviewers. (5) *Observations* [35] on the individual student or whole students in the classroom have been developed to assess engagement, which can be time-consuming for the administration and all kinds of

Table 1. Related works for engagement prediction with sensing data

Prediction	Data source	Participants	Data Sessions
Audience Engagement [95]	ACC data	10 children audience in art performance	not stated
Social Engagement [47]	EDA data (wristband)	Children during social interactions	51 sessions
Game Engagement [49]	EDA, PPG data	10 players in 6 mobile games in natual settings	not stated
Audience Engagement [41]	EDA, PPG data	10 attendess and 19 presentors in presentations	40 sessions
Student Engagement [1]	Video, audio data	25 university students in 5 classrooms	not stated
Student Engagement [63]	Video, heart rate data	23 university students in laboratory settings	not stated
Student Engagement [60]	EDA data (hand sensor)	17 undergraduate students in climate science classes	not stated
Student Engagement [92]	EDA data (hand sensor)	17 university students in learning environments	not stated
Emotional Engagement [27]	EDA data (wristband)	27 university students in 41 lectures over 3 weeks	197 sessions
Multidimensional Student Engagement (this work)	EDA, PPG, ST, ACC, CO2, Noise, etc.	23 high school students in 98 classes over 4 weeks	331 sesssions

academic settings need to be considered to get an accurate picture of student behaviour. The reliability of the observations can be doubtful as they only provide limited information about students.

All traditional methods for engagement measurement have strengths and limitations in different situations. Overall, traditional methods are usually time-consuming, and the quality of answers largely depends on the students, teachers, or executor. Recently, with the development of wearables and IoT sensors, some initial progress has been explored to measure student engagement with physiological signals which is more subjective and obtrusive to students.

2.2 Engagement Prediction with Sensing Technology

Sensing technologies are becoming prevalent to assess people's mental characteristics (e.g., engagement [27, 47–49, 95], mood [65, 93], stress [7, 93], personality [39]) and have provided an attractive alternative to traditional self-report surveys. Wang et al. [93] gathered students' mental health data such as mood and stress from self-report surveys in Dartmouth college. They also recorded students' activity data from passive sensors and found a significant correlation between the sensor data and mental health. Morshed et al. [65] predicted mood instability only using sensed data from mobile phones and wearable sensors for individuals in situated communities. Wang et al. [94] predicted human personality traits from passive sensing data from mobile phones using within-person variability features such as regularity index of physical activity, the circadian rhythm of location.

Physiological sensors and accelerometers have been explored to assess human's engagement (see Table 1), such as assessing audience engagement during the art performance, social engagement for children during the interaction with adults [47], emotional engagement for university students during lectures [27].

Ahuja et al. [1] built a classroom sensing system with commodity cameras. Students' and instructor's video and audio were captured for body segmentation and speech detection. Then, the students' engagement levels were analyzed based on their facial expressions and gestures. However, as reported from authors, this system would bring privacy concerns [54] when capturing audio and video data. Similarly, Hutt et al. [48] used commercial off-the-shelf eye-trackers to detect mind wandering for high school students and Monkaresi et al. [63] used heart rate and video-based estimation of facial expressions to predict the engagement of 23 university students during a structured writing activity in laboratory settings.

Only a few studies investigate student engagement in real-world settings [27, 37, 60, 92]. Mcneal et al. [60] used EDA hand-sensors to measure the engagement from 17 undergraduate students in climate science classes during a semester. They explored different teaching approaches on a subset of students and reported the statistical mean value of EDA traces. Contrast to their study, we collect a far more heterogeneous data set and novel features were proposed based on different physiological indices. Wang et al. [92] studied 17 university students' engagement in the distributed learning environment with EDA hand sensors, and found that EDA measurements were aligned with surveys. Different to our research, they only used a very simple question 'how much did you enjoy during the lecture' as the ground truth of students' engagement.

In recent years, researchers have started to explore different dimensions of engagement using physiological signals. Lascio et al. [27] predicted university students' emotional engagement from EDA sensors in lectures during 3-week data collection. While in our data collection, we build an in-class multidimensional (behavioural, emotional, cognitive) engagement sensing system including physiological responses (i.e., EDA, HRV, ST), physical movements (ACC) and indoor environmental sensors (i.e., CO_2, temperature, humidity, sound) for high school students. Furthermore, high school classes are very different from lectures at university in [27] (e.g., degree of freedom to choose courses, ability to schedule classes flexibly, requirements of class attendance, consistency of subjects between different schools), which may lead to very different multi-engagement distribution in high school classes. Another similar research, but for a different application, was proposed by Huynh et al. [49] who measured the engagement level of game players with multiple sensors. Though they agreed that user engagement includes three dimensions, they did not differentiate each dimension when predicting the engagement during the game. Nevertheless, in our study, we derive the different factors and most useful sensors contributing to the different dimensions of student learning engagement.

In summary, different from previous efforts, our work has several advantages: (1) we use far more heterogeneous data for engagement prediction (others only use EDA or heart rate data except [49]); (2) we propose and extract more meaningful features from physiological signals while [37, 60, 92] only use the simple average value of data; (3) to the best of knowledge, we are the first to predict the engagement for all three dimensions based on education research while previous studies either measure the simple general engagement, or a single dimension of engagement [27]), and derive the most useful sensors in differentiating each dimension of engagement; (5) we adopt real-world classroom settings and take the influence of environmental changes into account.

3 DATA COLLECTION

We conducted a field study in a private high school for 4 weeks in 2019. The data collection has been approved by the Human Research Ethics Committee at our University. We will then provide the details about participants, equipments to collect the data, and data collection procedures.

3.1 Participants

We recruited 23 students (13 females and 10 males, 15-17 years old) and 6 teachers (4 females and 2 males, 33-62 years old) in Year 10 (see Table 3). First, we gave an introduction to all Year 10 students and teachers, and distributed consent forms to them. Then, students or teachers who volunteered to participate returned the

Table 2. Room allocation for different class groups. Most classes belong to form group.

Room	Number of Students		
	Form	*Math*	*Language*
Room 1	10	7	13
Room 2	6	9	2
Room 3	7	7	5
Room 4	NaN	NaN	3

Table 3. Basic information for student and teacher participants.

Category	Students	Teachers
Total Number	23	6
Female	13	4
Male	10	2
Age	15-17 y.o.	33-62 y.o.

signed consent form from themselves and their (students') guardians. Once they signed forms, they were asked to complete the online background survey. Background information was collected at once including age, gender, how their engagement is affected by their thermal feelings, form class group, math class group, and language class group. There are 3 form classes in Year 10 and students are in form groups when having most classes (i.e., English, Global Politics, Science, Physical Education, Health/Sport). For Mathematics class, students are divided into 3 study groups and students in each group take classes in individual classrooms. For Language class, students have 4 different groups such as Japanese, French, Cultural Sustainability, and each group has classes in different classrooms too. Hence, the background information of different study groups helps us align students in different classrooms (see Figure 1(c)) at different times. We also recruited 3 Math teachers, 1 English teacher, 1 Japanese teacher and 1 Science teacher. Table 2 shows the details about room allocation for participants.

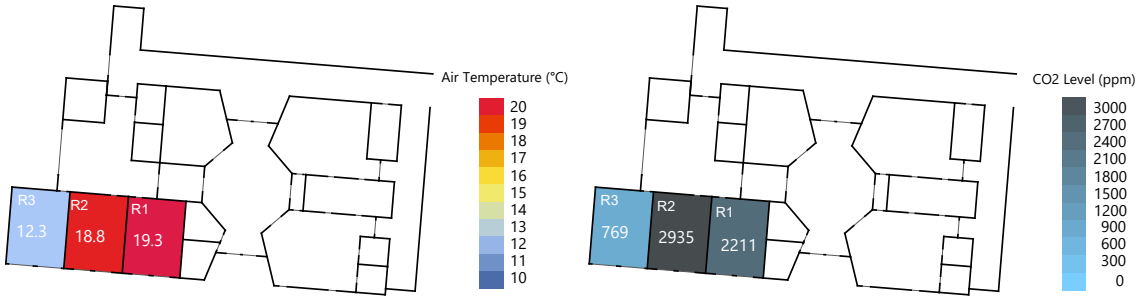
3.2 Collected Data

3.2.1 Physiological and Activity Data. During the school time, we asked participants to wear *Empatica E4* [1] wristbands as shown in Figure 1(a), first proposed in [40]. E4 wristband is a watch-like device with multiple sensors: electrodermal activity (EDA) sensor, photoplethysmography (PPG) sensor, 3-axis accelerometer (ACC), and optical thermometer. EDA depicts constantly fluctuating changes in skin electrical properties at 4 Hz. When the level of sweat increases, the conductivity of skin increases. PPG sensor measures the blood volume pulse (BVP) at 64 Hz, from which the inter-beat interval (IBI) and heart rate variability (HRV) can be derived. ACC records 3-axis acceleration in the range of [-2g, 2g] at 32Hz and captures motion-based activity. The optical

[1]Empatica E4 wristband: https://www.empatica.com/en-int/research/e4/

(a) Empatica E4 wristbands (b) Netatmo indoor weather station (c) Classroom for Year 10 students

Fig. 1. Devices and environments for collecting wearable and indoor data



(a) Temperature (b) CO_2 level

Fig. 2. Temperature and CO_2 data in R1, R2, R3 (room 1, room 2, room 3) at 11:00 am on 11 Sep 2019. Room 4 is not shown here as it is in another building.

thermometer reads peripheral skin temperature (ST) at 4 Hz. In the recording mode, E4 wristband can store 60 hours of data in the memory, and the battery can last for more than 32 hours. It is light-weight, comfortable and water-proof, thus especially suitable for continuous and unobtrusive monitoring of participants in our study.

3.2.2 Indoor Environmental Data. We collected indoor environmental data from the Netatmo Healthy Home Coach [2] - a smart indoor weather station - installed in the classrooms as shown in Figure 1(b) and Figure 1(c). The Netatmo station can collect indoor temperature (TEMP), humidity (HUMID), CO_2 and sound (SOUND) in every 5 minutes. Real-time data can be uploaded to the Cloud continuously through the Guest WiFi covered on the campus. Figure 2 shows the indoor temperature and CO_2 level in three rooms at 11:00 am on 11 Sep 2019. We can clearly see that the temperature of room 3 is only 12.3 ℃ and much lower than the comfortable warmth (18 ℃) defined by the World Health Organization's standard [69], which may negatively affect student learning in class [50]. Furthermore, CO2 levels in room 2 and room 3 are beyond 2000 ppm, which has been proved to have a negative influence on the student cognitive load in the classroom [45, 80]. Based on previous studies [72], students may become sleepy and inattentive during the class when the CO_2 level is too high.

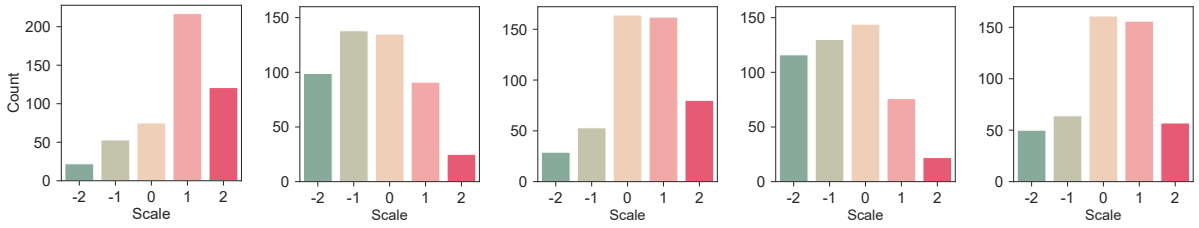
3.2.3 Ground Truth: Self-report Survey Instrument Data. In this study, we choose to use self-report survey to gather subjective measurements of students' in-class engagement. As discussed in Section 2, the self-report survey is the most common way to measure student engagement as they can reflect students' subjective perceptions explicitly. Instead, measures relying on experience sampling, teacher ratings, interviews or observations have been reported to be easily affected by the external factors. The questionnaire includes 5 items related to behavioural, emotional, and cognitive engagement of the validated *In-class Student Engagement Questionnaires* (ISEQ) [37], which has been proved to be effective for measuring multidimensional engagement compared to the traditional long survey. Similar to [27, 49], we slightly adapted survey questions from university lectures to high school class context to make the survey easier for students underage to understand. Moreover, for cognitive engagement measurement, we did not use the original question '*the activities really helped my learning of this topic*' in [37], considering that some classes in high school do not have in-class activities. Instead, we use the well-accepted item '*I asked myself questions to make sure I understood the class content*' [64], which is a good reflection of cognitive engagement. Table 4 shows the questionnaire used for measuring multidimensional student engagement in class, where item 1,3 and 5 assess the behavioural, emotional and cognitive engagement, item 2 and 4 indicate the behavioural and emotional disaffection [37, 87] .

[2]Netatmo Healthy Home Coach: https://www.netatmo.com/en-eu/aircare/homecoach

Table 4. Self-report items for measuring in-class engagement in online survey.

Questions (please describe your engagement in the last class)	Subscales
1. I paid attention in class.	Behavioural
2. I pretended to participate in class but actually not.	Behavioural (-)
3. I enjoyed learning new things in class.	Emotional
4. I felt discouraged when we worked on something.	Emotional (-)
5. I asked myself questions to make sure I understood the class content.	Cognitive

Note: (-) means the reversed score.



(a) Question 1 (b) Question 2 (c) Question 3 (d) Question 4 (e) Question 5

Fig. 3. Histograms of the Answers. The X axis shows the 5-Likert scale from -2 to 2 which means 'strongly disagree' to 'strongly agree'. The Y axis shows the number of the responses that fall into the specific scale.

In the questionnaire, each item [3] is rated with a 5-point Likert-scale from -2 to 2, which indicates 'strongly disagree', 'somewhat disagree', 'neither agree nor disagree', 'somewhat agree' and 'strongly agree'. Figure 3 shows the distribution of responses for each item from total 488 responses. The online self-report survey is constructed with the external tool named *Qualtrics* [4]. Participants were asked to complete the survey on the public tablets or their digital products with the given survey link generated by *Qualtrics*.

3.3 Procedure

Before the data collection, all wristbands were synchronized with the E4 Manager App from the same laptop to make sure the internal clocks are correct. 1 Netatmo weather station was installed and 1 tablet was put on the teacher desk in each classroom. Students were asked not to unplug the Netatmo stations during the data collection.

The first two weeks of data collection occurred in early September (winter in the southern hemisphere), and the next two weeks of data collection completed in November (spring in the southern hemisphere). We collected data from two different seasons to build a more robust engagement sensing system. As we know, different seasons usually result in different indoor environments (e.g., indoor temperature, humidity), which may affect students' sweat level (EDA, ST) and activity level (ACC, HRV). If we use the data from one season to build the engagement prediction model, the prediction performance can be greatly reduced in another season due to changes in activity, physiological, and environmental data. Before the data collection, 1 participant was chosen as the representative

[3]In the survey, participants were also asked to report their thermal feelings and mood using the Photographic Affect Meter (PAM) [75]. Nevertheless, this data was not considered in this paper.
[4]Qualtrics: https://www.qualtrics.com/au/

in each form class, for a total of 3 representatives. During the data collection, student participants were distributed with the same wristband (attached with the student ID label) from the representative at 8:50 before the first class started at 9:00. Then at the end of the school day (i.e., 15:35), the representative would remind student participants to hand in wristbands. Student participants were asked to wear the wristband on non-dominating hands and avoid pressing the button or performing any unnecessary movements during class. For teacher participants, they only need to wear the wristband during their classes.

On each school day, student participants were asked to complete the online surveys (either through the public tablets or their own digital devices) at 11:00, 13:25, 15:35 (right after the 2^{nd}, 4^{th}, 5^{th} class). The length of 2^{nd} and 4^{th} class can either be 40 minutes or 80 minutes on the different school day and the 5^{th} class always lasts for 80 minutes. From the class table for Year 10 students, they have the same class schedule on the 1^{st} week and 3^{rd} week, and another class schedule on the 2^{nd} and 4^{th} week. Each representative would remind student participants to complete online surveys on time. However, considering that it could be a burden for some participants to complete the survey 3 times a day, we did not urge students to complete the survey for ensuring the quality of survey responses. By the end of the 4^{th} week, we had received 488 valid responses in total and the response rate is 35.3%.

As a token of appreciation, the certificate of participation and four movie vouchers were provided to the participants during the 4-week data collection. It is worth noting that participation in this research project is voluntary. Participants are free to withdraw from the project at any stage if they change their minds. Besides, we anonymized all the participants to protect their privacy.

4 DATA PREPROCESSING

In this section, we first extract class periods based on students' accelerometer data using unsupervised time series segmentation method. Then we introduce the data cleaning process and data pre-processing technique for electrodermal activity, blood volume pulse, accelerometer data, and skin temperature data.

For data preparation, we only remain the data between 9:00 am to 15:35 pm, which corresponds to the start time of the first class and the end time of the last class. In addition, some students may have several data recording segments during the same day due to the unexpected closure and re-open of the wristband. We drop the data segments less than 15 seconds which is less helpful for extracting useful information. We also discard the data on Tuesday in the last week because students had trip travel and did not have classes on that day.

4.1 Class Period Segmentation

As described in Section 3, student participants wear wristbands all day along and teachers participants are only asked to wear the wristband at their classes. Participants report their engagement for the 2^{nd}, 4^{th}, 5^{th} classes of the day during recess time, lunchtime and before going home. Though the scheduled class start/end time is already known, teachers may start/finish the class a bit earlier or later than the scheduled time. The accurate class time is significant for wearable data analysis because participants may have very different physiological/movement patterns between in-class and after-class. For instance, increased activity level after class may lead to a higher value of EDA (due to the higher level of sweat) and variation of accelerometer data.

To get the exact class start/end time for meaningful data analysis, we segment the accelerometer data from student participants based on the assumption that students usually have different activity patterns before/after class. *Information-Gain based Temporal Segmentation* (IGTS) [26, 78] is applied on the ACC data to calculate the class start/end time. IGTS is an unsupervised segmentation technique, aims to find the transition times in human activities, which is suitable for dividing the boundary between in-class and out-class [78]. Topdown optimization is adopted in the ACC time-series segmentation. To calculate the class boundary, we choose the ACC time-series from 5 minutes before the class to 5 minutes after the class. Take calculating the actual class end time as an

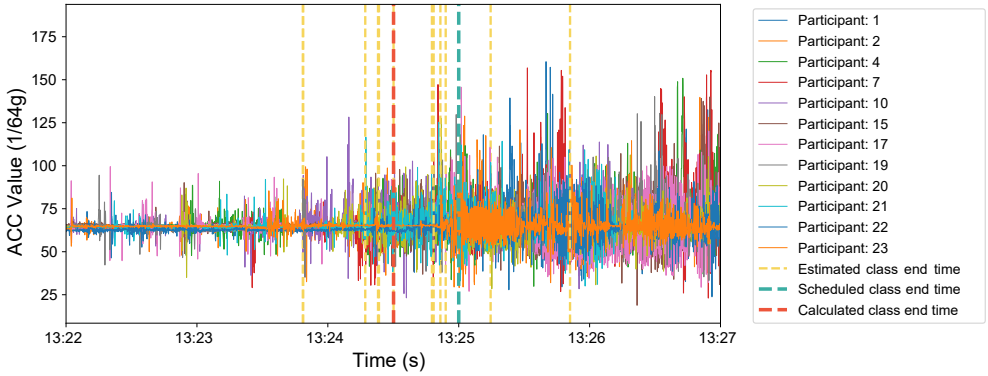


Fig. 4. Calculated class end time with ACC data from 12 student participants.

example, from Figure 4, there are 12 participants in a class and the scheduled class end time is 13:25 (green vertical dashed line). Applying IGTS on the ACC data, we can get 12 different estimated class end time from 12 ACC traces. Then, the median time is chosen as the calculated class end time (red vertical dashed line). That is to say, this class finishes early than the scheduled time. We apply IGTS on all the class data and extract the exact class start time and end time for the later data analysis.

4.2 Data Cleaning

Before pre-process the collected data, a data cleaning stage needs to be conducted to remove noises from wearable data. As describe in [7, 12, 27, 41, 47], there are several noises commonly happened in data collection from E4 wristband: (1) flat responses (i.e., 0 micro siemens) due to poor contact between the sensors and the skin. If the contact is not tight enough, the sensor will not measure anything; (2) abrupt signal drops due to the movement of the sensor (e.g., participant bumps the wristband onto the desk); (3) quantization errors. Since EDA sensor records data through the two electrodes, which is more susceptible to noises compared with ACC, PPG and ST sensor, we clean the data set mainly based on the quality of EDA data.

Firstly, we remove the data when students did not wear wristbands during the whole class or closed off the wristbands unintentionally during the class. Similar to [41], we then discard the signals containing a huge number of flat responses, abrupt signal drops and quantization error as suggested in [12, 47]. Finally, we discard the class data from the student who did not complete the survey. The data cleaning stage leaves us with 331 class data sessions. The final wearable data are gathered from 23 students and 6 teachers in 105 classes. 59 classes are short classes (mean = 39.15 minutes, STD = 1.15 minutes) and 46 classes are long classes with 2 periods (mean = 78.21 minutes, STD = 4.33 minutes).

The data cleaning stage brings to the elimination of 157 class data sessions due to the lack of survey data, which takes up to 32.17% of the total data with completed surveys. Though the number of eliminated data is considerable, the size of our collected data is comparable and even larger than the previous studies. For instance, Lascio et al. [27] used 197 EDA data sessions after a reduction up to 37%, Gashi et al. [41] used 40 presenter-audience EDA pairs with the elimination of 72 pairs. Hernandez et al. [47] used 51 data sessions with the elimination of 28% from the original data.

4.3 Data Pre-processing

The pre-processing procedure is crucial improve the quality of collected data. For EDA signals, we follow the same pre-processing steps as suggested in [27, 41, 47]. (1) Artifacts removal. To mitigate the influence of motion artifacts (MAs), we apply a median filter on EDA data with a 5-second window as in [27]. (2) Decomposition. EDA signal combines a tonic component and a phasic component [12, 14]. The tonic component varies slowly and reflects the general activity of sweat glands influenced by the body and environmental temperatures. The phasic component indicates rapid changes and related to the responses to internal and external stimuli. EDA signals are decomposed with cvxEDA approach [42] using convex optimization. (3) Normalization. The amplitude of the EDA signal varies a lot among different people [14] and thus limits the possibility of comparing the signal directly. We normalize the mixed, tonic and phasic EDA values similar to [41].

PPG data, also known as BVP, is provided by the E4 wristband. Similar to [49], we extract IBI signals by detecting the systolic peak of the heartbeat waveform signals from the raw PPG data (window size = 0.75 seconds). Linear interpolation is applied when the heartbeat intervals can not be detected successfully from the low-quality (e.g., motion artifacts) PPG signal. For the ACC data, we calculate the magnitude of 3-axis accelerations as $|a| = \sqrt{x^2 + y^2 + z^2}$. Then a median filter with 0.2 seconds is applied to the magnitude value. Finally, we apply a median filter on the ST data with 0.5 seconds.

5 FEATURE EXTRACTION

We use various sensing devices to infer multidimensional engagement level of high school students. Table 5 summarizes these features. Then, we introduce the computed features and discuss why we explore such sensors and features.

5.1 EDA-based Features

EDA is a common measure of autonomic nervous system activity, with a long history being used in psychological research [62]. Recently, EDA measurements have been increasingly explored in affective computing such as the detection of emotion [7, 16], depression [79] , and engagement [27, 47, 55]. From EDA data, we extract statistical features such as the standard deviation from EDA (mixed, tonic, phasic) data, which reflects the overall general arousal during the class [27]. As suggested in [27], we extract the number of arousing/arousing states, the ratio of arousing states, etc. to show the momentary engagement during the class. The similarity-based method such as Pearson Correlation Coefficient (PCC) [8] and Dynamic Time Wrapping Distance (DTW) [82] are used to evaluate the physiological synchrony [71] of the target student and teacher. Inspired by [95], we also propose some new features (marked with *) to compute physiological synchrony between the target student and the average values of other students, which has proven to be effective in Table 7.

5.2 HRV-based Features

HRV is controlled by the autonomic nervous system (ANS), which can be used to evaluate human emotional arousal and cognitive performance [2, 4, 19, 56, 68]. With the help of *HeartPy* [90] toolkit, we compute HRV features from IBI signals extracted from the raw PPG data. As suggested in [15, 61, 83], HRV features can be analyzed from time-domain and frequency domain. On the time-domain, we capture features such as the mean/standard deviation of RR intervals (Meani, SDNN) which estimates the overall HRV. We also extract features such as standard deviation/root mean square of successive RR interval differences (SDSD, RMSSD), number/percentage of successive interval pairs that differ larger than 20/50 ms (NN20, NN50, pNN20, pNN50), which describes the momentary change of HRV. On the frequency-domain where parameters are computed by applying Fast Fourier Transform (FFT) to the time series of RR intervals [83], we compute the absolute power of

Table 5. Description of the features computed for different sensors

Sensors	*Feature name*	*Description of features*
EDA	eda/tonic/phasic_avg	Average value for the raw, tonic, phasic data
	eda/tonic/phasic_std	Standard deviation for the raw, tonic, phasic data
	eda/tonic/phasic_n_p	Number of peaks for the raw, tonic, phasic data
	eda/tonic/phasic_a_p	Mean of peak amplitude for the raw, tonic, phasic data
	eda/tonic/phasic_auc	Area under the curve of the raw, tonic, phasic data
	num_arouse	Number of arousing moments during the class
	ratio_arouse	Ratio of arousing and unarousing moments
	level_k	Ratio of the number of level_k and the length of S_k
	eda/tonic/phasic_pcct	Pearson correlation coefficient with teacher
	eda/tonic/phasic_pccs*	Pearson correlation coefficient with average value of students
	eda/tonic/phasic_dtwt	Dynamic time wraping distance with teacher
	eda/tonic/phasic_dtws*	Dynamic time wraping distance with average value of students
PPG	hrv_bpm	Average beats per minutes
	hrv_meani	Overall mean of RR intervals (Meani)
	hrv_sdnn	Standard deviation of intervals (SDNN)
	hrv_lf_power	Absolute power of the low-frequency band (0.04–0.15 Hz)
	hrv_hf_power	Absolute power of the high-frequency band (0.15–0.4 Hz)
	hrv_ratio_lf_hf	Ratio of LF-to-HF power
	hrv_rmssd	Root mean square of successive RR interval differences
	hrv_sdsd	Standard deviation of successive RR interval differences
	hrv_pnn50	Percentage of successive interval pairs that differ >50 ms
	hrv_pnn20	Percentage of successive interval pairs that differ >20 ms
ACC	acc_avg	Average physical activity intensity during the class
	acc_std	Standard deviation of physical activity intensity in class
	acc_dtw_t	Dynamic time wraping distance with teacher
	acc_dtw_s*	Dynamic time wraping distance with average value of students
	acc_pcc_t	Pearson correlation coefficient with teacher
	acc_pcc_s*	Pearson correlation coefficient with average value of students
ST	sktemp_avg/max/min	Average/maximum/minimum value of skin temperature
CO2	mean/max/min_co2	Average/maximum/minimum value of CO2
TEMP	mean/max/min_temp	Average/maximum/minimum value of indoor temperature
HUMID	mean/max/min_co2	Average/maximum/minimum value of humidity
SOUND	mean/max/min_temp	Average/maximum/minimum value of sound

the low-frequency band (0.04-0.15 Hz) and high-frequency band (0.15-0.4 Hz). Besides, we compute the ratio of LF-to-HF power which reflects the overall balance of the ANS [67].

5.3 Accelerometer-based Features

Student behaviour can be inferred from ACC data, which helps us know more about student participation (e.g., team activities) and engagement level in class [95]. For ACC data, we extract features such as the average physical

activity and standard deviation, which describes the statistical characteristics of the student movement during the class. Inspired by [95], we propose the movement synchrony features such as the DTW/PCC between the target student and the average values of the other students.

5.4 Other Features

Student learning engagement has been found to be affected by the thermal comfort level of students in the classrooms [50], which is influenced by many factors such as indoor temperature, humidity, skin temperature, sound, CO_2 level, etc [25, 38, 69, 76]. Therefore, statistical features are calculated for indoor temperature, CO2, sound and humidity, as the overall estimate of the indoor environment during a class. For ST data, statistical features are extracted to estimate the general arousal of student engagement. According to [51], when CO_2 level is higher than 1000 ppm, occupants may complain about the drowsiness and poor air, and when CO_2 level is higher than 2000 ppm, occupants will feel sleepy, headaches and lose attention. Therefore, the above features are selected to study student engagement.

6 PREDICTION PIPELINE

Although engagement prediction is usually regarded as a classification problem, where engagement level can be divided into two or three categories [27, 49] based on specific thresholds, it is not a good practice to determine people's psychological characteristics using classification [39]. In this paper, we choose regression rather than classification for multidimensional engagement prediction. In order to predict multidimensional engagement scores of students, we set up a regression-based pipeline as described below.

Engagement Score: We assign each student a score for each item in the self-report survey. To achieve this, we first reverse the responses in item 2 and item 4, as shown in Table 4. Then, we calculated a score based on the average of the 5-point Likert scale for each dimension of engagement and the overall engagement. Then we rescale the calculated score to 1 to 5, representing the engagement level being low to high. Figure 5 shows the calculated overall engagement score for 23 student participants. To save space, we do not display box plots of the distribution of the single-dimensional engagement score.

Regressors: We adopt LightGBM Regressor [52, 84] to predict self-reported multidimensional engagement scores. As one of the most powerful prediction models, LightGBM is an ensemble method combining a set of weak predictors (i.e., regression trees) to make accurate and reliable predictions. It builds the regression tree vertically (leaf-wise) while other algorithms grow trees horizontally (level-wise). It will choose the leaf with max delta loss to grow. When growing the same leaves, LightGBM algorithm can reduce more loss than other tree-based algorithms such as GBRT [28].

Validation: It is natural to use cross-validation to train and test prediction models when we are not in a data-rich situation. The purpose of cross-validation is to estimate the unbiased generalization performance of the prediction model. However, when using the test set for both model selection (hyperparameter tuning) and model estimation, the test data may be overfitted, and the optimistic bias may occur in the model estimation. Therefore, we adopt the nested cross-validation approach [66] with inner loop cross-validation nested in outer loop cross-validation. The inner loop is used for hyperparameter tuning and feature selection, while the outer loop is responsible for evaluating the performance on the test set. In the outer loop, similar to the previous human-centred research [27, 47], we first divide the data into n groups, where n represents the number of participants, i.e., n=23. Each group contains the data for only one participant. Then we apply *k-fold cross-validation* [96] (k=5) and on all student groups. Specifically, data from the same student (group) will not appear in the training and test sets at the same time. In the inner loop, the remaining data groups are split into L (L=3) folds, where each fold serves as a validation set in turn. Then we train (grid search) the hyperparameters on the training set, evaluate them on the validation set, and select the best parameter settings based on the performance recordings over

Table 6. Prediction performance for emotional, cognitive, behavioural, and overall engagement with all sensing data

Dimension	*MAE*				*RMSE*			
	LGBM.	LR.	Average	Random	LGBM.	LR.	Average	Random
Emotional	**0.675**	0.714	0.747	1.059	**0.851**	0.878	0.928	1.326
Cognitive	**0.906**	0.921	0.977	1.288	**1.113**	1.128	1.176	1.658
Behavioural	**0.783**	0.811	0.871	1.235	**0.960**	0.980	1.135	1.540
Overall	**0.602**	0.614	0.641	0.891	**0.753**	0.769	0.792	1.125

L folds. We use the importance vector generated from LightGBM to reduce the feature dimensionality, which calculates feature importance automatically by averaging the number of times a specific feature used for splitting a branch. Higher values indicate higher feature importance. Top-10 features are selected as the new input features to the LightGBM regressor. The heuristic of choosing 10 features is we find that the prediction error is lowest under this threshold in the experiment.

Similar to [27, 94], we also perform leave-one-subject-out (LOSO) [36] validation to evaluate the impact of data from individual participant on the overall prediction error. For both k-fold and LOSO validation approaches, we calculate the average performance score (i.e., MAE and RMSE) of the regressor in each iteration.

Baselines and Metrics: We compare the proposed engagement prediction model with three baselines. The first baseline is the standard linear regressor [81], one of the most widely used regression models. The second baseline takes the average score of each dimension of engagement. The third baseline randomly generates a sample from the distribution of engagement scores and regards it as a predicted value. Similar random baselines have been widely used in previous ubiquitous computing studies such as [27, 94]. To evaluate the prediction performance of the proposed model, we use the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [18] metrics.

7 RESULTS AND DISCUSSION

In this section, we conduct extensive experiments to evaluate the prediction performance of *n-Gage*. We answer the first research question '*Can we measure the multiple dimensions of high school student's learning engagement including emotional, behavioural and cognitive engagement in high schools with sensing data in the wild?*' in Section 7.1. We answer the second research question '*Can we derive the activity, physiological, and environmental factors contributing to the different dimensions of student learning engagement? If yes, which sensors are the most useful in differentiating each dimension of the engagement?*' in Section 7.2. We also study how different settings can help improve the performance of *n-Gage*. Unless otherwise stated, the prediction models are built with LightGBM regressors using all sensors and evaluated by k-fold nested cross-validation by default.

7.1 Overall Prediction Results

We first evaluate the overall prediction results for *n-Gage* with all sensors available. Table 6 displays MAE and RMSE scores of *n-Gage*'s engagement regression in different dimensions. In particular, the overall engagement is calculated by the average of engagement scores from all questions related to the engagement, which is commonly used in previous engagement studies [27, 34, 49]. From Table 6, we can see that in terms of MAE and RMSE, *n-Gage* achieves higher prediction performance for all dimensions of engagement than all baselines, demonstrating its potential for multidimensional engagement prediction.

Notably, among each dimension of engagement, *n-Gage* works best on predicting emotional engagement. The emotional engagement regression model obtain 0.675 of MAE and 0.851 of RMSE, which is lower than 0.384

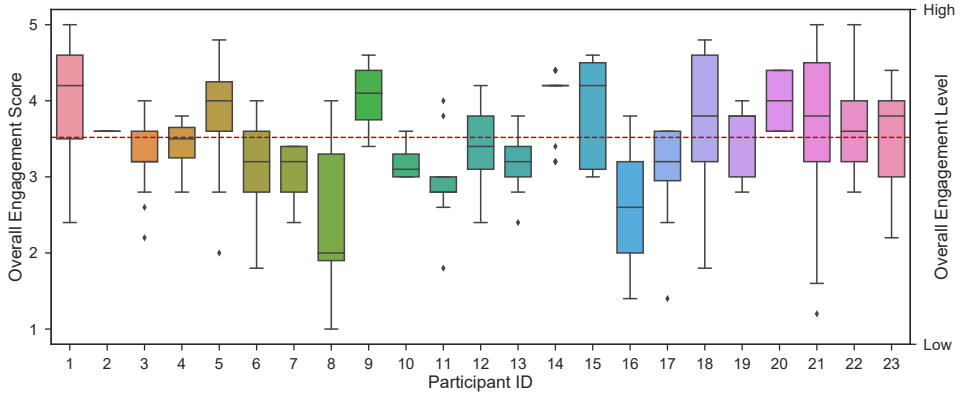


Fig. 5. Box plot of the overall engagement scores for 23 student participants. The red dashed line represents the average score for all participants. The participant ID shown in the figure is randomly generated to maintain the privacy of participants.

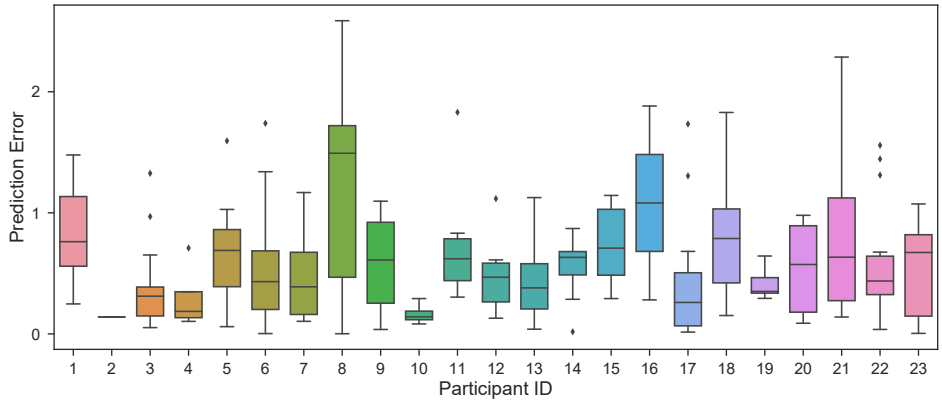


Fig. 6. Prediction error for overall engagement scores for 23 student participants

(36.26%) and 0.475 (35.82%) of the random baseline. The reasons why *n-Gage* predicts emotional engagement best are possibly two-fold: (1) compared with cognitive and behavioural engagement, emotional engagement is most suitable for evaluation through self-report surveys [34], resulting to a more realistic and stable student emotional engagement measurement (ground truth). (2) emotional engagement is more easily detected by sensors (e.g., EDA and PPG) as it reflects the degree of emotional arousal, thereby producing fluctuations in physiological signals [7, 27, 53].

Although the MAE of cognitive engagement regression is higher than other models, it is still lower than random baseline of 0.382 (29.66%) in MAE and 0.545 (32.87%) in RMSE. The possible reason is that cognitive engagement is more challenging to be assessed by the wearable and indoor sensors than electroencephalography (EEG) sensors [9]. By contrast, *n-Gage* has the lowest prediction error of 0.602 in MAE and 0.753 in RMSE in overall engagement assessment. According to the education research [34, 35], although the multidimensional concept of engagement has been well accepted, the definitions of three dimensions of engagement vary with

Table 7. The most influential features on multidimensional engagement.

Engagement	*Association*	*Most influential features*
Emotional Engagement	(+)	acc_pcc_s, tonic_a_p, eda_pcc_s
	(-)	**acc_avg***, **sktemp_avg***, eda_dtw_t
Cognitive Engagement	(+)	**intemp_min***, level_1, hrv_ratio_lf_hf
	(-)	**acc_pcc_s***, co2_max, acc_std
Behavioural Engagement	(+)	acc_std, acc_pcc_s, eda_pcc_avg
	(-)	**sktemp_avg***, **acc_pcc_t***, acc_dtw_t
Overall Engagement	(+)	level_1, tonic_a_p, intemp_max
	(-)	**acc_dtw_t***, sktemp_avg, acc_avg

* indicates p-value < 0.01.

considerable overlap across components. Therefore, the overall engagement is easier to be evaluated and predicted than the single-dimensional engagement.

We also compare the prediction results with when a standard linear regressor is learned. From Table 6, the linear regression model has much higher prediction performance than both average and random baseline models (e.g., 31.09% lower than random baseline model in MAE for overall engagement prediction), indicating the effectiveness of extracted features in engagement prediction. However, the performance of linear regressors is not comparable to the LightGBM in all dimensions. This is because LightGBM has a good ability to capture non-linear feature-target relationships which is more flexible than simple linear regressors. To summarize, we believe that the performance of *n-Gage* is benefited from both the extracted features and powerful non-linear mapping provided by the LightGBM.

We then discuss the impact of data from the individual participant on the overall prediction error. We train and test the regressors using the LOSO validation approach which enables us to evaluate the ability of models to accurately predict a new participant not included in the training set. Figure 6 shows the boxplot of absolute prediction error per participant. Interestingly, each participant has a very different error distribution. For instance, participants 8 and 16 have the highest median value (1.492 and 1.082) and standard deviation (0.801 and 1.132) of prediction errors. From Figure 5, we observe that both participants have a much lower engagement level than the others. Since the regression model is built on the data from all the other participants, it does not work well when the participant (testing set) has a different distribution from the training set. The potential solution is to build participant-wise or groupwise prediction models, as introduced in [70]. In conclusion, we believe the prediction errors come from both the specific participants and overall prediction bias. We will further investigate this issue in future research.

7.2 Impact of Sensor Combinations

We will explore the physiological, activity and environmental factors contributing to the different dimensions of student engagement. We compute the Pearson Correlation Coefficient (PCC) between the extracted features and multiple dimensions of engagement, and then list the three most influential features in Table 7. We find many EDA features related to the peaks of tonic EDA signals and physiological synchrony are related to the multidimensional engagement. In previous research, EDA features are generally considered as a good indicator of physiological arousal (e.g., emotional and cognitive states) [12, 23], which have been explored in the detection of engagement [27, 47, 55]. For the HRV features (e.g., 'hrv_ratio_lf_hf'), they are shown to be correlated with cognitive engagement as HRV is an autonomically dependent variable and has been used to predict student

Table 8. Summary of the Prediction performance of multidimensional engagement using different sensor combinations. $\mathcal{X}_1$ indicates all the wearable data including EDA, HRV, ACC and ST data, and $\mathcal{X}_2$ means the indoor environmental data including CO_2 and temperature data.

Data source	*MAE/RMSE*			
	Emotional	*Cognitive*	*Behavioural*	*Overall*
EDA	0.697/0.877	0.948/1.149	0.851/1.019	0.637/0.800
HRV	0.714/0.901	0.940/1.140	0.833/1.002	0.659/0.812
EDA+HRV	0.699/0.875	0.949/1.151	0.841/0.989	0.621/0.783
EDA+ACC	0.679/0.860	0.914/1.124	0.816/0.987	0.626/0.789
HRV+ACC	0.691/0.875	0.910/1.125	0.809/0.979	0.641/0.796
EDA+HRV+ACC	0.679/0.860	0.909/1.122	0.800/0.965	0.620/0.778
$\mathcal{X}_1$ *	**0.673**/0.851	0.910/1.126	0.811/0.980	0.619/0.775
$\mathcal{X}_1 + \mathcal{X}_2$ * *(all)*	0.675/0.851	**0.906/1.113**	**0.783/0.960**	**0.602/0.753**

* indicates the proposed combination of features for engagement prediction.

engagement in [63]. Similar to EDA and HRV features, we notice that the average skin temperature ('sktemp_avg') are negatively correlated with engagement, as ST reflects the sympathetic nervous activity and attention states [3] which has been used for mind-wandering prediction [11] and stress detection [46].

For activity factors, it is interesting to find that many ACC features are highly correlated with engagement. Accelerometer is a popular and powerful sensor for quantifying human behavioural patterns [39, 93]. ACC features have been utilised to sense audience engagement using interpersonal movement synchrony [95]. In the experiment, we observe that the average physical intensity during class is highly negatively correlated with emotional engagement. This leads us to believe that when students are negatively engaged, they tend to perform more physical movements in the class. As for environmental factors, we find that the maximal CO_2 level is negatively associated with cognitive engagement, while the indoor temperature is positively associated with engagement. This may be because CO_2 has a negative impact on people's cognitive load [45, 80], and then affects student cognitive engagement. This result highlights the need to ventilate the classroom timely to keep students engaged. Interestingly, we notice that the maximal indoor temperature in the class is positively correlated with overall engagement. One possible explanation is that during the data collection period (winter and spring), the indoor temperature is low and moderately higher indoor temperature makes students feel thermally comfortable [69] and therefore more engaged in learning [50].

Then we investigate the most useful sensors in predicting each dimension of student engagement and explore the performance of *n-Gage* when only a set of sensors available. In this research, we use E4 wristbands and Netatmo indoor weather stations for student engagement assessment. However, when other schools want to generalize the system for automatic engagement measurement, it is likely that only a few sensors available considering the types of wearables and installation of indoor weather stations. In this experiment, we use different combinations of sensors as shown in Table 8 to train the regressors, where $\mathcal{X}_1$ indicates all the wearable sensors including EDA, HRV, ACC and ST, and $\mathcal{X}_2$ represents all the environmental sensors containing CO_2, TEMP, HUMID and SOUND sensors. Besides, we predict student engagement using only EDA as in [27], single PPG (HRV) as in [37], and EDA+HRV as in [49]. Since accelerometers are naturally available in wearables and have been used for engagement measurement [95], we add ACC to the above sensor combinations for the first time. Then we utilise all wearable sensors and indoor sensors for more accurate engagement prediction.

Table 9. Multidimensional Engagement Regression Result for Different Subjects

Subject	MAE/RMSE			
	Emotional	Cognitive	Behavioural	Overall
Maths	0.686/0.841	0.841/0.965	0.750/0.891	0.603/0.738
English	0.609/0.779	0.893/1.010	0.694/0.819	0.510/0.629
Language	0.645/0.814	0.829/0.903	0.799/0.900	0.593/0.758
Science	0.646/0.829	0.895/0.941	0.758/0.856	0.575/0.720
Politics	0.674/0.835	0.947/1.057	0.660/0.731	0.525/0.671
Average	0.652/0.820	0.881/0.975	0.732/0.839	0.561/0.703

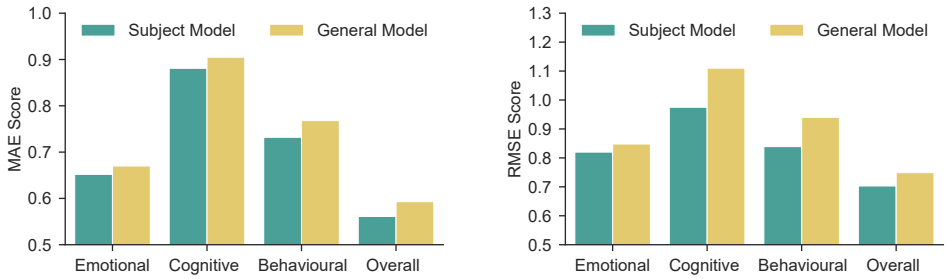
For each sensor combination, we use nested cross-validation to train and test the regressors as described in Section 6, to achieve optimal feature selection and parameter tuning. Table 8 displays the regression result with different sensor combinations. Different combinations are useful for different dimensions of engagement. For instance, a single EDA sensor works well for emotional engagement prediction while less useful in predicting behavioural engagement unless involving ACC together. This is reasonable because EDA is a reflection of emotional arousal, while ACC is capable of quantifying human behavioural patterns [39, 93]. On the other hand, the combination of EDA and HRV sensors has similar prediction performance compared to using a single EDA sensor, which is consistent with the fact that not many HRV features are highly correlated with engagement. When there is no EDA sensor (especially in commercial off-the-shelf smart wristbands), the HRV+ACC combination can achieve similar prediction performance on cognitive and behavioural engagement compared to EDA+HRV+ACC.

Meanwhile, it can be observed that the combination of all wearable sensors ($\mathcal{X}_1$) has the lowest prediction error for emotional engagement. When considering wearable sensors ($\mathcal{X}_1$) with indoor sensors ($\mathcal{X}_2$), *n-Gage* can achieve the best performance on the behavioural, cognitive and overall engagement, and has similar prediction performance in emotional engagement with $\mathcal{X}_1$. The underlying reason is that CO_2 and indoor temperature mainly affect students' cognition load and behavioural patterns. For example, students may lose attention (related to behavioural engagement), sleepy (related to cognitive engagement) [51] during class when the CO_2 level is too high (e.g., larger than 2000 ppm), but this does not necessarily mean that students do not like the class (related to emotional engagement). The above results illustrate the importance of taking indoor environmental changes into account for student engagement prediction and creating the optimal environment to keep students engaged in class.

7.3 Impact of Class Subjects

Now, we investigate whether considering different school subjects could improve the prediction performance of *n-Gage*. Our assumption here is that different subjects may lead to different learning requirements, thinking styles and emotional preferences. Then, student engagement levels and physiological status may be affected accordingly.

To validate this hypothesis, we establish regression models for each subject (i.e., Language, Maths, Science, English, PE, Politics, Health, Chapel) to isolate differences in class subjects and engagement assessment. Table 9 summarizes MAE and RMSE scores of the regressors over different subjects. We do not consider the Health, Chapel and PE classes because the number of survey responses are limited (less than 30) in those classes which may affect the prediction performance. We also compare the average prediction performance of 5 regression models (i.e., Maths, English, Language, Science, Politics) with the general regressor model in Figure 7. The results



(a) MAE Score (b) RMSE Score

Fig. 7. Prediction performance for the average subject model and general model

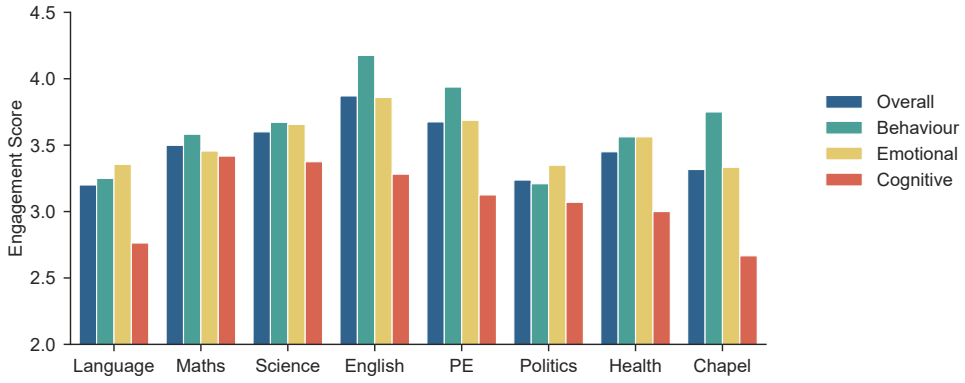


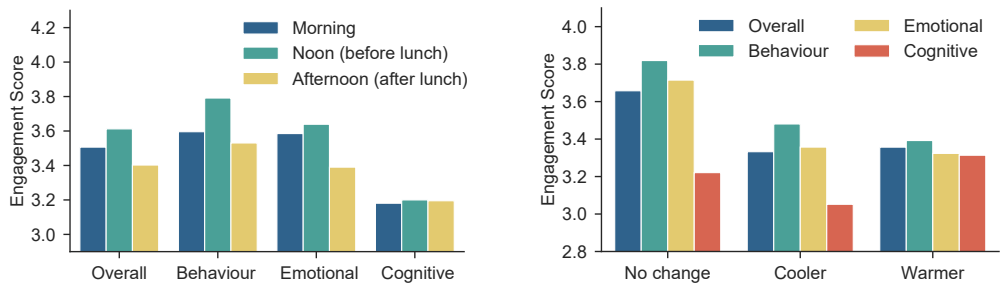
Fig. 8. Engagement scores on different subjects

indicate that, compared with building the general regression model including all subjects, building regression models by school subjects can significantly improve the prediction performance.

To better understand the underlying cause behind the improved regression performance, we review the self-reported engagement scores. Figure 8 shows that students have very different multidimensional engagement scores among different subjects. For instance, while students have the highest behavioural and emotional engagement score in English class, they have the highest cognitive engagement score in Maths class. The possible reason is that students enjoy English classes most and thus like to follow the rules from English teachers. Due to the fact that Math know-how is cumulative and usually contains complex concepts, students may tend to put more effort to comprehend the contents in Maths class, thus leading to a high cognition engagement score. Overall, these observations serve as evidence that building models for each subject can lead to significantly improved prediction performance.

7.4 Discussion

We have shown that it is possible to infer multidimensional student engagement by using multiple wearable and environmental sensors. Meantime, we will present the following interesting discussion points.



(a) Class time

(b) Thermal preference

Fig. 9. Engagement scores with different class time and thermal comfort

- **Engagement and class time**. A preliminary study is conducted to investigate the correlation between self-reported student engagement and class time during the school day. Figure 9(a) shows the average engagement scores for the different class time (morning, noon and afternoon). Overall, we observe that classes in the noon show higher engagement levels in all dimensions. Classes in the afternoon (after lunch) have the lowest engagement score, especially in the behavioural and emotional dimensions. Particularly, it is interesting to notice that students have a much higher behavioural and emotional engagement level than the cognition level despite the time of the classes. These observations provide directions for further research in maximizing student engagement by a more reasonable arrangement of class schedule according to the nature of each course.
- **Engagement and thermal comfort**. In the background survey, most students agree that *'When I am engaged in class, I could get distracted when the room is too hot or too cold'* (see Section 3.2.2). As another investigative point, Figure 9(b) shows the relationship between self-reported engagement and thermal preference (i.e., warmer, cooler, no change) [25] of the students in class. The results show that students who feel the room thermally comfortable have a higher overall engagement level compared to other groups. In particular, students who prefer a cooler environment usually have the lowest cognition engagement. This reminds us that creating the right thermally comfortable environment is necessary to improve student engagement in class, especially considering the individual differences in thermal sensation [30].
- **Real-time measurement**. Students' engagement level during a class may vary with the learning content and teaching style. Real-time anonymous engagement tracking can provide teachers with student engagement level and help teachers understand the impact of different teaching contents on student engagement, thereby better adjusting teaching speed and teaching methods. However, the challenge is how to obtain the fine-grained ground truth of student engagement multiple times during the class without disturbing students' studying. One potential approach is ecological momentary assessment (EMA) [53] which repeatedly prompts students to report their engagement level. Though EMA is usually considered a good method of *in situ* data collection, if students need to answer EMA, they may be disturbed and distracted in class. Overall, ground truth data collection is challenging and more reliable methods need to be investigated.

8 IMPLICATIONS AND LIMITATIONS

This research addresses the possibility of automatically predicting students' in-class emotional, behavioural and cognitive engagement using wearable and indoor sensing technology, which provides opportunities for the future

design of feedback systems in the classroom. The feedback system has the potential to benefit both teachers and students.

Teacher plays an important role in influencing student engagement [34]. With the feedback from students after each class, teachers can evaluate, and, if necessary, adapt or change teaching strategies (e.g., increase time for student thinking, allow students time to write, assign reporters for small groups [89]) for creating the right learning climate to keep students engaged [44]. For instance, when teachers focus more on academics and fail to create a positive social learning environment, students are likely to be emotionally disengaged and worried about making mistakes. Contrarily, when teachers focus more on the social dimension and neglect the intellectual dimension, students possibly experience low cognitive engagement for learning [34, 88]. With such a feedback system, teachers can observe multidimensional student engagement and create the intellectually challenging and socially supportive learning environment.

Further, if this system is deployed, using *n-Gage*, teachers can take timely measures to improve learning experience for students, such as planning learning schedules, re-engaging students with the low engagement, and ventilating the room to let the fresh air in. While overcoming student disengagement is complicated, we do believe teachers can benefit from the engagement feedback of students after every class instead of few times in a term [17, 27], contributing to higher student achievements and protecting students from dropping out of school [34].

Students wearing wristbands are able to self-track their multidimensional in-class engagement, which positively influences academic achievements and is usually regarded as the predictor of learning outcomes [20, 34, 58]. Being conscious of in-class engagement is an effective *quantified-self* [29, 77] approach to promote self-regulation and reflective learning [10] for students. Once students are aware of how much effort they are putting into learning, they can work towards their personal goals by optimizing their study practices and learning strategies (e.g., practice active listening and thinking, make study plans for different subjects) [6, 29]. Additional strategies such as gamification [13, 24] can also be deployed along with *n-Gage* measurements.

For real-world deployments, the feedback system can still work when only a subset of sensors available (see Section 7.2). For instance, when there are no indoor sensors installed, wearable sensors can be used for accurate engagement prediction especially for the emotional engagement. The system can also allow more sensors to be integrated in the future when becomes available.

The current studies have some limitations that needed to be addressed in future research. Firstly, collecting data from more student participants in the same class may bring new opportunities for data analysis. There are 59 Year 10 students in total, but only 23 students voluntarily become participants and wear wristbands. Compared to students who did not participate, participants may share some similar personality traits and have higher potential to engage in class most of the time.

Secondly, we agree that collecting the ground truth of student engagement is challenging because we need to find a compromise between taking long psychological surveys for more accurate measurement and enabling students to complete surveys faster without affecting their study or rest. Therefore, a more robust way of evaluating multidimensional student engagement needs to be investigated in the future.

Thirdly, the quality of survey responses varies. Online surveys are conducted 3 times a day, and the total response rate is 35.3%. Since completing surveys multiple times a day may become a burden, students are likely to answer the questions unseriously. Therefore, in this study, we only encourage rather than urge them to complete the survey, which to a certain extent guarantees the quality of responses. Figure 10 shows the survey completion time for all responses from participants. Most participants complete the survey in 30 to 50 seconds, but some participants complete the survey in less than 15 seconds. Though the survey completion time may be affected by many factors and varies from person to person, it is still one of the indicators of response quality [57]. In future research, it will be interesting to explore patterns from survey completion time data and assign appropriate weights to survey response for more accurate prediction of student engagement.

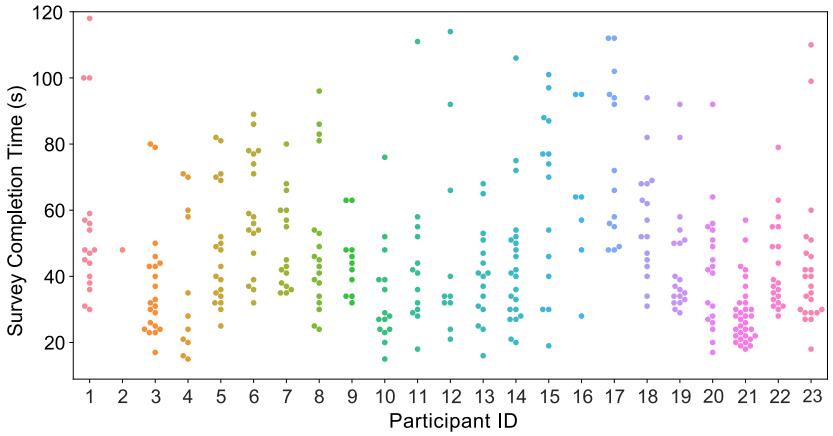


Fig. 10. Survey completion time for different participants. Each point represents the survey completion time for one response.

During the in-situ data collection, the data recorded by the wristband is not always continuous. For many reasons, we face a considerable loss of data: (1) on each school day, an average of 2 to 4 participants got sick leave and cannot wear a wristband; (2) 6 participants went abroad to a study program on the second week of data collection; (3) students were curious about the wristbands, especially in the first few days, and they pressed the button again and again out of curiosity. Some students accidentally closed their wristbands, so their data was lost for hours or even the whole day.

Though significant efforts have been made to make the maximal use of the collected data, 32.17% traces must to be removed from the analysis due to the loss of survey data, the incomplete data during the class, the presence of long-time of flat responses, artifacts and quantization errors as discussed in Section 4. Despite the fact that we have cleaned and pre-processed wearable data to eliminate noises, collecting physiological data in the wild still faces huge challenges, especially for young students. In our research, one of the main noise is from the poor contact between the sensors and skin, which can be fixed by tightening the wrist strap to the skin. However, this will also increase awareness of wristband during class, resulting in student in-class disengagement and even more motion artifacts.

9 CONCLUSION AND FUTURE WORK

In this research, we propose *n-Gage*, an engagement sensing system that can capture students' physiological responses, physical movements, and environmental changes to infer multidimensional engagement (behavioural, emotional and cognitive engagement) level in class. We evaluate the system by combining weather station data and wearable data collected from 23 Year 10 students and 6 teachers over 144 classes in 4 weeks in a high school. Some new features are proposed to characterize different aspects of student engagement. Extensive experiment results show that *n-Gage* can predict student behavioural, emotional and cognitive engagement score (1 is the lowest score and 5 is the highest score) with an average MAE of 0.788 and RMSE of 0.975. We further demonstrate the most influential features and how different sensor combinations/school subjects affect student engagement. Finally, we show some interesting findings that the maximal CO_2 level is highly negatively correlated with student cognitive engagement; class time (morning, noon and afternoon) and thermal preference (warmer, cooler or no change) may affect the level of student engagement, which provides beneficial insights for educators and school managers to improve student learning engagement in high school.

Though not perfect, we believe that *n-Gage* is still a very promising first-step towards multidimensional in-class engagement tracking for students. As a contribution, *n-Gage* can indicate the future design of feedback system, assisting students and teachers in a variety of ways (e.g., promoting students' self-regulation and reflective learning, helping teachers create a right learning climate for students). In the future, we plan to involve more participants of different ages from different schools in data collection. The expansion of the dataset will help us get better and more precise results. Also, we hope to investigate more factors that may affect students' multi-engagement, such as personality, mood and behavioural habits.

ACKNOWLEDGMENTS

This research is supported by the Australian Government through the Australian Research Council's Linkage Projects funding scheme (project LP150100246). This paper is also a contribution to the IEA EBC Annex 79.

REFERENCES

[1] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical classroom sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 71.

[2] Nutan D Ahuja, Amit K Agarwal, Ninad M Mahajan, Naresh H Mehta, and Hatim N Kapadia. 2003. GSR and HRV: Its Application in Clinical Diagnosis. In *16th IEEE Symposium Computer-Based Medical Systems, 2003. Proceedings.* IEEE, 279–283.

[3] John L Andreassi. 2010. *Psychophysiology: Human Behavior and Physiological Response.* Psychology Press.

[4] Bradley M Appelhans and Linda J Luecken. 2006. Heart Rate Variability as an Index of Regulated Emotional Responding. *Review of general psychology* 10, 3 (2006), 229–240.

[5] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. 2006. Measuring Cognitive and Psychological Engagement: Validation of the Student Engagement Instrument. *Journal of school psychology* 44, 5 (2006), 427–445.

[6] Kimberly E Arnold, Brandon Karcher, Casey V Wright, and James McKay. 2017. Student Empowerment, Awareness, and Self-regulation Through a Quantified-self Student Tool. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference.* 526–527.

[7] Jorn Bakker, Mykola Pechenizkiy, and Natalia Sidorova. 2011. What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data. In *2011 IEEE 11th international conference on data mining workshops.* IEEE, 573–580.

[8] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson Correlation Coefficient. In *Noise reduction in speech processing.* Springer, 1–4.

[9] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. 2007. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. *Aviation, space, and environmental medicine* 78, 5 (2007), B231–B244.

[10] Paul Black and Dylan Wiliam. 2009. Developing the Theory of Formative Assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)* 21, 1 (2009), 5.

[11] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. 2014. Automated Physiological-based Detection of Mind Wandering During Learning. In *International Conference on Intelligent Tutoring Systems.* Springer, 55–60.

[12] Wolfram Boucsein. 2012. Electrodermal Activity: Springer Science & Business Media. *Broek, EL vd, Schut, MH, Westerink, JHDM, Herk, J. v., & Tuinenbreijer, K* (2012).

[13] Patrick Buckley and Elaine Doyle. 2016. Gamification and student motivation. *Interactive learning environments* 24, 6 (2016), 1162–1175.

[14] John T Cacioppo, Louis G Tassinary, and Gary Berntson. 2007. *Handbook of Psychophysiology.* Cambridge University Press.

[15] A John Camm, Marek Malik, J Thomas Bigger, Günter Breithardt, Sergio Cerutti, Richard J Cohen, Philippe Coumel, Ernest L Fallen, Harold L Kennedy, RE Kleiger, et al. 1996. Heart Rate Variability: Standards of Measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. (1996).

[16] Filipe Canento, Ana Fred, Hugo Silva, Hugo Gamboa, and André Lourenço. 2011. Multimodal Biosignal Sensor Data Handling for Emotion Recognition. In *SENSORS, 2011 IEEE.* IEEE, 647–650.

[17] Steven Cantrell and Thomas J Kane. 2013. Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-year Study. *MET Project Research Paper* (2013).

[18] Tianfeng Chai and Roland R Draxler. 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?–Arguments Against Avoiding RMSE in the Literature. *Geoscientific model development* 7, 3 (2014), 1247–1250.

[19] Kwang-Ho Choi, Junbeom Kim, O Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji-Eun Park. 2017. Is Heart Rate Variability (HRV) an Adequate Tool for Evaluating Human Emotions?–A Focus on the Use of the International Affective Picture System (IAPS). *Psychiatry research* 251 (2017), 192–196.

[20] James Patrick Connell, Margaret Beale Spencer, and J Lawrence Aber. 1994. Educational Risk and Resilience in African-American Youth: Context, Self, Action, and Outcomes in School. *Child development* 65, 2 (1994), 493–506.
[21] Lyn Corno and Ellen B Mandinach. 1983. The Role of Cognitive Engagement in Classroom Learning and Motivation. *Educational psychologist* 18, 2 (1983), 88–108.
[22] National Research Council et al. 2003. *Engaging Schools: Fostering High School Students' Motivation to Learn*. National Academies Press.
[23] Hugo D Critchley. 2002. Electrodermal Responses: What Happens in the Brain. *The Neuroscientist* 8, 2 (2002), 132–142.
[24] Marguerite Cronk. 2012. Using Gamification to Increase Student Engagement and Participation in Class Discussion. In *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 311–315.
[25] Richard J De Dear and Gail S Brager. 2002. Thermal Comfort in Naturally Ventilated Buildings: Revisions to ASHRAE Standard 55. *Energy and buildings* 34, 6 (2002), 549–561.
[26] Shohreh Deldari, Jonathan Liono, Flora D Salim, and Daniel V Smith. 2019. Inferring Work Routines and Behavior Deviations with Life-logging Sensor Data. (2019).
[27] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 103.
[28] Jane Elith, John R Leathwick, and Trevor Hastie. 2008. A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology* 77, 4 (2008), 802–813.
[29] Rebecca Eynon. 2015. The Quantified Self for Learning: Critical Questions for Education.
[30] Asma Ahmad Farhan, Krishna Pattipati, Bing Wang, and Peter Luh. 2015. Predicting Individual Thermal Comfort Using Machine Learning Algorithms. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 708–713.
[31] Jeremy D Finn. 1989. Withdrawing from School. *Review of educational research* 59, 2 (1989), 117–142.
[32] Jeremy D Finn, Gina M Pannozzo, and Kristin E Voelkl. 1995. Disruptive and Inattentive-withdrawn Behavior and Achievement Among Fourth Graders. *The Elementary School Journal* 95, 5 (1995), 421–434.
[33] Jeremy D Finn and Donald A Rock. 1997. Academic Success Among Students at Risk for School Failure. *Journal of applied psychology* 82, 2 (1997), 221.
[34] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of educational research* 74, 1 (2004), 59–109.
[35] Jennifer A Fredricks and Wendy McColskey. 2012. The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments. In *Handbook of research on student engagement*. Springer, 763–782.
[36] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics New York.
[37] Kathryn A Fuller, Nilushi S Karunaratne, Som Naidu, Betty Exintaris, Jennifer L Short, Michael D Wolcott, Scott Singleton, and Paul J White. 2018. Development of a self-report Instrument for Measuring In-class Student Engagement Reveals that Pretending to Engage is a Significant Unrecognized Problem. *PloS one* 13, 10 (2018), e0205828.
[38] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, Jun Zhai, Klaus David, and Flora D Salim. 2020. Transfer Learning for Thermal Comfort Prediction in Multiple Cities. *arXiv preprint arXiv:2004.14382* (2020).
[39] Nan Gao, Wei Shao, and Flora D Salim. 2019. Predicting Personality Traits From Physical Activity Intensity. *Computer* 52, 7 (2019), 47–56.
[40] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3—A wearable Wireless Multi-sensor Device for Real-time Computerized Biofeedback and Data Acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 39–42.
[41] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 13.
[42] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.
[43] James E Groccia. 2018. What is student engagement? *New Directions for Teaching and Learning* 2018, 154 (2018), 11–20.
[44] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of educational research* 77, 1 (2007), 81–112.
[45] Ulla Haverinen-Shaughnessy, DJ Moschandreas, and RJ Shaughnessy. 2011. Association Between Substandard Classroom Ventilation Rates and Students' Academic Achievement. *Indoor air* 21, 2 (2011), 121–131.
[46] Katherine A Herborn, James L Graves, Paul Jerem, Neil P Evans, Ruedi Nager, Dominic J McCafferty, and Dorothy EF McKeegan. 2015. Skin Temperature Reveals the Intensity of Acute Stress. *Physiology & behavior* 152 (2015), 225–230.
[47] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. 2014. Using Electrodermal Activity to Recognize Ease of Engagement in Children During Social Interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive*

and Ubiquitous Computing. ACM, 307–317.
[48] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D'Mello. 2019. Automated Gaze-based Mind Wandering Detection During Computerized Learning in Classrooms. *User Modeling and User-Adapted Interaction* 29, 4 (2019), 821–867.
[49] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 13.
[50] Han Jiang, Matthew Iandoli, Steven Van Dessel, Shichao Liu, and Jacob Whitehill. 2019. Measuring Students' Thermal Comfort and Its Impact on Learning. *Educational Data Mining* (2019).
[51] Kane. 2020. What are Safe Levels of CO and CO2 in Rooms? https://www.kane.co.uk/knowledge-centre/what-are-safe-levels-of-co-and-co2-in-rooms Accessed 2020-07-08.
[52] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*. 3146–3154.
[53] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-wild Stress in Pregnant Mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–22.
[54] Mikkel B Kjærgaard, Omid Ardakanian, Salvatore Carlucci, Bing Dong, Steven K Firth, Nan Gao, Gesche Margarethe Huebner, Ardeshir Mahdavi, Mohammad Saiedur Rahaman, Flora D Salim, et al. 2020. Current Practices and Infrastructure for Open Data based Research on Occupant-centric Design and Operation of Buildings. *Building and Environment* (2020), 106848.
[55] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. 2011. Love, Hate, Arousal and Engagement: Exploring Audience Responses to Performing Arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1845–1854.
[56] Antonio Luque-Casado, Mikel Zabala, Esther Morales, Manuel Mateo-March, and Daniel Sanabria. 2013. Cognitive Performance and Heart Rate Variability: the Influence of Fitness Level. *PloS one* 8, 2 (2013), e56935.
[57] Neil Malhotra. 2008. Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly* 72, 5 (2008), 914–934.
[58] Helen M Marks. 2000. Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High school years. *American educational research journal* 37, 1 (2000), 153–184.
[59] Jonathan Martin and Amada Torres. 2016. What is Student Engagement and Why is it Important. *Retrieved May* 4 (2016), 2018.
[60] Karen S McNeal, Jacob M Spry, Ritayan Mitra, and Jamie L Tipton. 2014. Measuring Student Engagement, Knowledge, and Perceptions of Climate Change in an Introductory Environmental Geology Course. *Journal of Geoscience Education* 62, 4 (2014), 655–667.
[61] Bruce Mehler, Bryan Reimer, and Ying Wang. 2011. A Comparison of Heart Rate and Heart Rate Variability Indices in Distinguishing Single-Task Driving and Driving Under Secondary Cognitive Workload. *Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011* (2011). https://doi.org/10.17077/drivingassessment.1451
[62] Wendy Berry Mendes. 2009. Assessing Autonomic Nervous System Activity. *Methods in social neuroscience* (2009), 118–147.
[63] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated Detection of Engagement Using Video-based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing* 8, 1 (2016), 15–28.
[64] KA Moore and L Lippman. 2005. Conceptualizing and Measuring Indicators of positive Development: What do Children Need to Flourish.
[65] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 75.
[66] Andreas C Müller, Sarah Guido, et al. 2016. *Introduction to Machine Learning with Python: A Guide for data scientists*. " O'Reilly Media, Inc.".
[67] H Nagendra, Vinod Kumar, and Shaktidev Mukherjee. 2015. Cognitive Behavior Evaluation Based on Physiological Parameters Among Young Healthy Subjects with Yoga as Intervention. *Computational and mathematical methods in medicine* 2015 (2015).
[68] Peter Nickel and Friedhelm Nachreiner. 2003. Sensitivity and Diagnosticity of the 0.1-Hz component of Heart Rate Variability as an Indicator of Mental Workload. *Human factors* 45, 4 (2003), 575–590.
[69] World Health Organization et al. 2007. Housing, Energy, and Thermal Comfort. *A review of* 10 (2007).
[70] Manoranjan Pal and Premananda Bharati. 2019. *Applications of Regression Techniques*. Springer.
[71] Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. 2017. Interpersonal Autonomic Physiology: A Systematic Review of the Literature. *Personality and Social Psychology Review* 21, 2 (2017), 99–141.
[72] Wargocki Pawel, Alí Porras-Salazar José, and William P. Bahnfleth. 2017. Quantitative Relationships Between Classroom CO2 Concentration and Learning in Elementary Schools. *8th AIVC Conference "Ventilating healthy low-energy buildings"* (2017).
[73] Richard Pflanzer and W McMullen. 2013. Galvanic Skin Response and the Polygraph. *BIOPAC Systems, Inc. Retrieved* 5 (2013).
[74] Paul R Pintrich and Elisabeth V De Groot. 1990. Motivational and Self-regulated Learning Components of Classroom Academic Performance. *Journal of educational psychology* 82, 1 (1990), 33.

[75] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: A Photographic Affect Meter for Frequent, in Situ Measurement of Affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734.
[76] Mohammad Saiedur Rahaman, Jonathan Liono, Yongli Ren, Jeffrey Chan, Shaw Kudo, Tim Rawling, and Flora D Salim. 2020. An Ambient-Physical System to Infer Concentration in Open-plan Workplace. *IEEE Internet of Things Journal* (2020).
[77] Verónica Rivera-Pelayo, Valentin Zacharias, Lars Müller, and Simone Braun. 2012. Applying Quantified Self Approaches to Support Reflective Learning. In *Proceedings of the 2nd international conference on learning analytics and knowledge*. 111–114.
[78] Amin Sadri, Yongli Ren, and Flora D Salim. 2017. Information Gain-based Metric for Recognizing Transitions in Human Activities. *Pervasive and Mobile Computing* 38 (2017), 92–109.
[79] Marco Sarchiapone, Carla Gramaglia, Miriam Iosue, Vladimir Carli, Laura Mandelli, Alessandro Serretti, Debora Marangon, and Patrizia Zeppegno. 2018. The Association Between Electrodermal Activity (EDA), Depression and Suicidal Behaviour: A Systematic Review and Narrative Synthesis. *BMC psychiatry* 18, 1 (2018), 22.
[80] Usha Satish, Mark J Mendell, Krishnamurthy Shekhar, Toshifumi Hotchi, Douglas Sullivan, Siegfried Streufert, and William J Fisk. 2012. Is CO2 an Indoor Pollutant? Direct Effects of Low-to-moderate CO2 Cncentrations on Human Decision-making Performance. *Environmental health perspectives* 120, 12 (2012), 1671–1677.
[81] George AF Seber and Alan J Lee. 2012. *Linear Regression Analysis*. Vol. 329. John Wiley & Sons.
[82] Pavel Senin. 2008. Dynamic Time Warping Algorithm Review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855, 1-23 (2008), 40.
[83] Fred Shaffer and JP Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in public health* 5 (2017), 258.
[84] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jeffrey Chan, Xinhong Hei, Bradley Feest, and Flora D Salim. 2019. Flight Delay Prediction using Airport Situational Awareness Map. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 432–435.
[85] David J Shernoff, Mihaly Csikszentmihalyi, Barbara Schneider, and Elisa Steele Shernoff. 2014. Student Engagement in High School Classrooms from the Perspective of Flow Theory. In *Applications of flow in human development and education*. Springer, 475–494.
[86] Ellen Skinner, Carrie Furrer, Gwen Marchand, and Thomas Kindermann. 2008. Engagement and Disaffection in the Classroom: Part of a Larger Motivational Dynamic? *Journal of educational psychology* 100, 4 (2008), 765.
[87] Ellen A Skinner, Thomas A Kindermann, and Carrie J Furrer. 2009. A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. *Educational and Psychological Measurement* 69, 3 (2009), 493–525.
[88] Deborah Stipek. 2002. Good Instruction is Motivating. In *Development of achievement motivation*. Elsevier, 309–332.
[89] Kimberly D Tanner. 2013. Structure Matters: Twenty-one Teaching Strategies to Promote Student Engagement and Cultivate Classroom Equity. *CBE—Life Sciences Education* 12, 3 (2013), 322–331.
[90] Paul van Gent, Haneen Farah, Nicole van Nes, and Bart van Arem. 2019. HeartPy: A Novel Heart Rate Algorithm for the Analysis of Noisy Signals. *Transportation research part F: traffic psychology and behaviour* 66 (2019), 368–378.
[91] Jorina von Zimmermann, Staci Vicary, Matthias Sperling, Guido Orgs, and Daniel C Richardson. 2018. The Choreography of Group Affiliation. *Topics in Cognitive Science* 10, 1 (2018), 80–94.
[92] Chen Wang and Pablo Cesar. 2015. Physiological Measurement on Students' Engagement in a Distributed Learning Environment. In *PhyCS*. 149–156.
[93] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 3–14.
[94] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. 2018. Sensing Behavioral Change over Time: Using Within-person Variability Features from Mobile Sensing to Predict Personality Traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 141.
[95] Jamie A Ward, Daniel Richardson, Guido Orgs, Kelly Hunter, and Antonia Hamilton. 2018. Sensing Interpersonal Synchrony Between Actors and Autistic Children in Theatre Using Wrist-worn Accelerometers. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 148–155.
[96] Trevor S Wiens, Brenda C Dale, Mark S Boyce, and G Peter Kershaw. 2008. Three Way K-fold Cross-validation of Resource Selection Functions. *Ecological Modelling* 212, 3-4 (2008), 244–255.