Using wearable physiological sensors for affect-aware intelligent tutoring systems

Alqahtani, Fehaid; Katsigiannis, Stamos; Ramzan, Naeem

Published in:
IEEE Sensors Journal

DOI:
10.1109/JSEN.2020.3023886

E-pub ahead of print: 14/09/2020

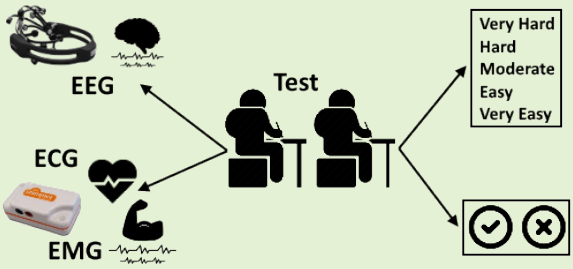
Document Version
Peer reviewed version

Alqahtani, F., Katsigiannis, S., & Ramzan, N. (2020). Using wearable physiological sensors for affect-aware intelligent tutoring systems. *IEEE Sensors Journal*. https://doi.org/10.1109/JSEN.2020.3023886

Using wearable physiological sensors for affect-aware Intelligent Tutoring Systems

Fehaid Alqahtani, Stamos Katsigiannis, *Member, IEEE*, and Naeem Ramzan, *Senior Member, IEEE*

***Abstract*— Intelligent Tutoring Systems (ITS) have shown great potential in enhancing the learning process by being able to adapt to the learner's knowledge level, abilities, and difficulties. An aspect that can affect the learning process but is not taken into consideration by traditional ITS is the affective state of the learner. In this work, we propose the use of physiological signals and machine learning for the task of detecting a learner's affective state during test taking. To this end, wearable physiological sensors were used to record electroencephalography (EEG), electrocardiography (ECG), and electromyography (EMG) signals from 27 individuals while participating in a computerised English language test. Features extracted from the acquired signals were used in order to train machine learning models for the prediction of the self-reported difficulty level of the test's questions, as well as for the prediction of whether the questions would be answered correctly. Supervised classification experiments showed that there is a relation between the acquired signals and the examined tasks, reaching a classification F1-score of 74.21% for the prediction of the self-reported question difficulty level, and a classification F1-score of 59.14% for predicting whether a question was answered correctly. The acquired results demonstrate the potential of the examined approach for enhancing ITS with information relating to the affective state of the learners.**



***Index Terms*— Affective computing, ECG, EEG, EMG, Intelligent Tutoring Systems (ITS), Machine Learning, Physiological Signals**

I. Introduction

INTELLIGENT Tutoring Systems (ITS) are systems designed to assist in the learning process by providing immediate and customised feedback and/or instructions to their users, requiring minimal to no input by instructors after the design of the learning material. Their advantage over traditional learning systems lies in their ability to adapt to the abilities, knowledge, and needs of individual learners, thus providing a learning experience tailored to the needs of each user [1]. The main target and focus of ITS is to facilitate the process of learning. Learning can be defined as an internal process of change, resulting from the learner's personal experience. Also, it can be defined as the acquisition or addition of something new, which involves any variation or modification previously acquired. Teachers guide students during the learning process and must perceive the students' needs in order to improve the teaching. However, in group tutoring environments, one-on-one time dedicated by instructors to each student decreases considerably. To address this issue, some researchers propose the use of ITS.

The efficiency of such systems over the traditional learning process is still not universally proven. Some studies disprove the notion of objective superiority of human tutoring and show that through properly refined software algorithms, ITS are on par with human tutoring and constitute an effective learning solution [2]. Contrary to that conclusion, other research works consider that traditional ITS are still not as effective as one-on-one tutoring [1] due to the lack of a human instructor that can understand the affective state of the students/learners and respond accordingly. Advocates of this view propose the integration of affective computing technologies into ITS, since they consider that emotions play an essential role in the learning process and human thinking [1], [3]–[5], thus learning environments have to consider this fact in order to be successful. Their proposed Affective Tutoring Systems enhance ITS by being able to adapt not only to the knowledge level and the abilities of the learner, but also to their affective state during the learning process [6]–[9].

Affective computing is "*computing that relates to, arises from, or deliberately influences emotions*" [10]. The field of affective computing focuses on the interpretation and recognition of emotions and the affective state of the user, as well as on the required methods for such interpretation according

Manuscript submitted on ...

F. Alqahtani was with the University of the West of Scotland, School of Computing, Engineering and Physical Sciences, Paisley, PA1 2BE, UK. He is now with King Fahad Naval Academy, Computer Science Department, Jubail 35512, Kingdom of Saudi Arabia (e-mail: Fehaid.Alqahtani@uws.ac.uk).

S. Katsigiannis was with the University of the West of Scotland, School of Computing, Engineering and Physical Sciences, Paisley, PA1 2BE, UK. He is now with Durham University, Dept. of Computer Science, Durham, DH1 3LE, UK (e-mail: stamos.katsigiannis@durham.ac.uk).

N. Ramzan is with the University of the West of Scotland, School of Computing, Engineering and Physical Sciences, Paisley, PA1 2BE, UK (e-mail: Naeem.Ramzan@uws.ac.uk).

to the user needs [11]. The human brain is a very complex system [12] and there have been diverse attempts to observe, understand, and model behaviour resulting from human response to stimuli, based on either empirical understandings developed by psychology [13], [14] or by the use of medical imaging (e.g. functional Magnetic Resonance Imaging - fMRI [15]) or using a variety of bio-signals, including physiological signals [16], [17].

Physiological signals are signals that are produced by the physiological process of human beings (e.g. heart beat, brain activity, muscle activity) that are affected in response to the central and the peripheral nervous system (CNS, PNS) of the human body. Research has shown that such signals, e.g. electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), etc., contain information related to the affective state of an individual, thus they can be exploited for the task of affect recognition [16]. Various studies [16]–[20] have established a relation between physiological signals and the Arousal and Valence dimensions of a felt emotion, as defined in Russel's *Circumplex Model of Affect* [21]. These studies focused on the use of pattern recognition and machine learning in order to extract spatial and spectral features from physiological signals and use them to train machine learning models in order to map the acquired features to the respective emotional state (in terms of Valence and Arousal). Combined with the use of wearable wireless non-invasive sensors for physiological signals that have become available in recent years (e.g. EEG [22], [23], ECG [24], [25], etc), such emotion recognition techniques can be integrated into ITS in order to allow adaptation according to the affective state of the learner.

In this work, we expanded our preliminary work [26], [27] and studied the potential use of EEG, ECG, and EMG physiological signals for detecting the affective state of users participating in a computerised English language test. The aim of the study was to examine whether features extracted by the aforementioned physiological signals are related to the difficulty level of each test question as perceived by the test takers during the test, and whether they can be used as an indicator of the success of the test taker in answering a test question. Establishing such links to the physiological signals could potentially benefit an ITS by allowing it to adapt the difficulty of the questions or focus on the questions that the users find more difficult. Achieving this would not be easy by a traditional ITS, since usually some answers are produced by guessing and other answers are produced by analysis using memorised facts and careful thinking. Furthermore, the lack of a human tutor that would be able to understand the affective state of the learners and adapt the learning process accordingly, further impedes the effectiveness of a traditional ITS as a learning medium. The use of physiological signals for detecting the affective state of the learners could potentially substitute this role of a human tutor in ITS. To study this proposal, portable, wearable, wireless EEG, ECG, and EMG sensors were used in order to record physiological signals from 27 individuals that took the computerised English language test. Features extracted from the acquired recordings were then used to train machine learning models for the prediction of the self-reported difficulty level of each question and of whether a question was answered correctly. Both single-subject and multi-subject models were trained, achieving a classification F1-score of 74.21% for the prediction of the self-reported question difficulty level, and a classification F1-score of 59.14% for predicting whether a question was answered correctly.

The rest of this work consists of five sections. Section II provides a brief literature review of the fields of ITS and affect recognition. The experimental protocol followed is described in Section III, while Section IV provides a thorough description of the proposed data analysis approach. Results are then presented and discussed in Section V, whereas conclusions are finally drawn in Section VI.

II. Background

ITS have been extensively studied within both theoretical and practical scenarios. Nevertheless, there is no single unified set of characteristics and methodologies that defines an ITS, with the term ITS being a broad descriptor that entails a large number of techniques that can be applied towards the same goal, i.e. facilitating the learning process. In a recent review of 50 different ITS focusing on various fields of study [28], the researchers concluded that improvements in the learning outcome when using an ITS depend upon the nature and the quality of the used ITS. A review [29] covering ITS studies between 1997 and 2010 concluded that ITS help more with improving course-specific evaluation than generalised testing and that while ITS have been shown to be effective in many areas compared to traditional tutoring, they are still not advanced enough to completely replace standard practices. An additional problem identified by Nye [30] is that current ITS designs are well-suited to developed countries but their use is challenging for the developing world. Furthermore, in the same review, it is stated that published ITS research suffers from a selection bias, since most systems are implemented and tested in environments that are well-suited for their use.

The typical architecture of ITS consists of four modules that are presented with various names within the literature [31]: The first is the expert module which includes the knowledge that the system is designed to pass to the learners (domain knowledge), as well as the techniques for analysing the learners' activities during the learning process. The second is the student diagnosis or student module which is built by gathering and updating information about the learner during the learning process, such as responses, behaviours, level of knowledge, learning style, etc. The third is the instruction/tutor/pedagogical module which detects knowledge deficiencies and focuses on applying specific strategies or teaching methods for compensating the knowledge deficit and the difficulty in learning. The methods used by this module include among others adaptive feedback, hints, recommendations, and navigation of the learning path. The fourth module consists of the user interface which is used for the communication and interaction between the ITS and the learner.

One aspect of the learning process that is not taken into consideration by conventional ITS is the affective/emotional state of the learner. Research on learning has shown that emotions

affect learning, with negative emotions impairing learning and positive emotions contributing to learning achievement [3]. Andres *et al.* [32] studied the patterns of educationally-relevant affective states within the context of an ITS and concluded that boredom is a powerful indicator of students' knowledge but not always indicative of learning, while delight is more weakly associated with knowledge. The affective state of learners engaged with ITS was also the focus of the Bosch and D'Mello study [33], which showed that engagement, confusion, frustration, boredom, and curiosity were the most frequent affective states, while confusion + frustration and curiosity + engagement were identified as two frequently co-occurring pairs of states.

Based on the effect of the affective state on the learning process, various researchers proposed the introduction of an affect detection mechanism into ITS [1], [34]. Kort *et al.* [35] proposed a model that conceptualised the impact of emotion in learning and attempted to recognise the cognitive-emotive state (affective state) of users within the context of a computerised learning companion. Ben Ammar *et al.* [3] proposed the use of facial expressions for the observation and detection of students' affective behaviour and responses. The detected responses were utilised by an ITS to adapt its teaching strategy and stimulate cooperative learning among learners. Facial features in combination with neural networks were also used by Zatarain-Cabada *et al.* [36] to allow a mobile device-based (smart-phones, tablets) ITS to adapt according to the emotional state of the user. Barrón-Estrada *et al.* [37] combined facial and voice features for the detection of emotion by an ITS designed and implemented within a social network.

The field of affective computing has been instrumental in the success and realisation of the aforementioned research works. One of the most important tasks in the field of affective computing is the task of affect recognition. Multiple works have studied the use of physiological signals to achieve this task. Zeng *et al.* [38] have conducted an extensive survey on emotion recognition techniques relying on various stimuli for affect elicitation. Gunes *et al.* [39] conducted a survey on continuous affect detection, while more recently Marechal *et al.* [40] provided a survey on multimodal methods for emotion recognition. Most works in the literature focus on the use of physiological signals for the extraction of features that are used to train machine learning models for the detection of affective states, in terms of the continuous Valence and Arousal scale.

In two extensively cited works, Soleymani *et al.* [18] studied the use of peripheral physiological signals in combination with eye gaze data and Support Vector Machines (SVM) for affect recognition when using film clips as the stimulus, while for the same task, Koelstra *et al.* [16] examined the use of EEG and peripheral physiological signals along the Naive-Bayes classifier. Various studies have used the Soleymani *et al.* [18] and the Koelstra *et al.* [16] datasets for evaluating affect recognition methods. For example, Arnau-González *et al.* [41] evaluated connectivity-based and channel-based EEG features, Mert and Akan [42] evaluated empirical mode decomposition (EMD) and its multivariate extension (MEMD) for emotion recognition, and Pereira *et al.* [43] examined the relation between EEG signal duration and the effectiveness of emotion recognition methods.

Various physiological signal modalities and sensors have been examined for the task of emotion recognition via physiological signals. Katsigiannis and Ramzan [17] compiled an emotion recognition dataset using wireless, portable, low-cost devices using film clips as stimulus and examined commonly used features along the SVM classifier, while Abadi *et al.* [44] examined additional modalities such as magnetoencephalography (MEG), electrooculography (hEOG), and near-infra-red (NIR) imaging along the Naive-Bayes classifier while using music videos as stimulus. The use of low-cost wireless devices for physiological signal acquisition was also explored by Correa *et al.* [45] along with facial and full body videos for the task of emotion recognition. In another recent work [46], the authors examined the use of affect recognition techniques for affect detection during the task of human-horse interaction, in order to facilitate equine-assisted therapy. The use of wearable low-cost EEG, ECG, and EMG devices along with various machine learning techniques demonstrated significant potential for affect recognition, in terms of Valence and Arousal, under the examined task.

Some works that employ affective computing techniques via physiological signal analysis in the context of ITS have been proposed in the literature. Physiological signals (EEG, ECG, GSR, etc.) have mostly been used for the task of emotion recognition during interaction with an ITS [47]–[53]. EEG signals have also been used in order to predict when a user makes a mistake while interacting in a dynamic learning environment (DLE) [54], or to determine mental engagement during problem solving tasks [55]. In another study, the use of physiological signals has also been proposed for quasi real-time adaptation in ITS [56]. Looking through the available literature, it is evident that the use of physiological signals in the field of ITS is focused mainly on emotion recognition and ITS adaptation. Based on this and to the best of the authors' knowledge, our preliminary work [26], [27] and this work are the first that attempted to establish the relation between physiological signals and the self-assessed difficulty level of answered test questions, as well as the success rate in answering the examined questions.

Considering the previous work in the literature as well as the practical requirements of a study related to ITS, wearable and wireless physiological signal sensors were used in this work for physiological signal acquisition.

III. Experimental Protocol

In order to study the relation between the affective state of an individual while interacting with an ITS, participants were recruited and were asked to complete a computerised English language test while physiological signals were recorded. Participants were also asked to provide feedback in relation to the difficulty of each question. Furthermore, the number of successfully answered questions was used in order to group the participants according to their English language level. Approval to conduct this study, including the acquisition and publication of anonymised data, was granted by the Ethics Committee of the University of the West of Scotland.

Fig. 1: Emotiv EPOC+ EEG sensor [22]

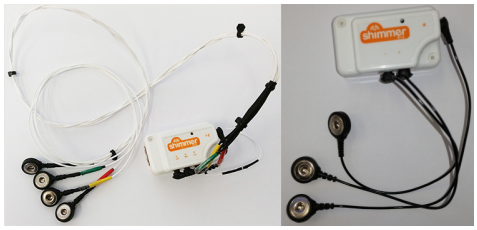


Fig. 2: SHIMMER ECG/EMG sensors [24]

A. Data acquisition

EEG, ECG, and EMG signals were recorded during each session of this study. EEG is used in order to monitor the electrical activity of the brain at specific locations, ECG provides a recording of the electrical activity of the heart, whereas EMG provides a recording of the electrical activity of the muscles that the EMG electrodes are attached to. Portable wireless lightweight sensors were selected for the acquisition of all the signals in order to minimise intrusiveness and discomfort to the participants of the study, thus minimising any bias stemming from the presence of the equipment. Furthermore, a laptop computer was used for the recording of the transmitted signals and for monitoring their quality. 14-channel EEG signals were captured at a 256 Hz sampling rate using an Emotiv EPOC+ wireless EEG headset [22] that utilises 16 gold-plated contact sensors fixed on flexible plastic arms, as shown in Fig. 1. To use the device, the contact sensors of the headset were placed against the head of the user at locations that closely align with the AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, M1 and M2 locations. The contact sensors located at M1 and M2 are used as reference and the rest 14 contact sensors are used for EEG data recording. The Emotiv EPOC+ headset is a relatively low cost EEG device that has been widely used in affective computing research (e.g. [17], [45], [46]), with the validity of its captured data being verified in [22] and [57]. The Emotiv EPOC+ EEG headset was selected due to its practicality. Recording EEG signals is an arduous and complex task, with medical-grade EEG equipment being bulky, requiring the use of electrode caps and conductive gel, having multiple cables attached to the cap that restrict movement, and requiring specialised technicians to operate. The Emotiv EPOC+ headset offers a practical solution to all these issues, at the cost of having lower resolution compared to medical-grade devices. Nevertheless, various works in the field of affective computing have demonstrated its efficiency (e.g. [17], [45], [46]). Two SHIMMER^{TM} v2 wireless sensors [24] (Fig. 2) were used for the acquisition of the ECG and EMG signals at a 256 Hz sampling rate. The ECG sensor utilised four standard electrodes positioned on both lower ribs and clavicle, while the EMG sensor utilised three standard electrodes positioned on the upper trapezius muscles.

B. Experimental setting

The experiment took place within a quiet office with no external noises and distractions. After initially explaining the experimental procedure, participants were given the opportunity to ask questions. Then, after signing a consent form, they were asked to sit in front of a computer. The supervising researcher proceeded with attaching the physiological sensors to the participants or with guiding them to attach them themselves when the electrodes had to be attached on the skin below the clothes. Participants were instructed to avoid excessive body and head movement during the experiment in order to reduce motion artefacts in the recorded signals. It must be noted that participants were discouraged from consuming caffeine or drugs before the experiment in order to avoid any related effects on the physiological signals. Then, the experiment started after correct signal transmission and acquisition was verified.

The experiment consisted of completing a computerised English language test on the laptop computer using a mouse for answering the questions. Twenty questions were selected from the Oxford Quick Placement Test (QPT) [58] for this experiment. The Oxford QPT contains 40 questions of varying difficulty and is designed for measuring the English language knowledge of the test takers, as well as for placing them, as accurately and reliably as possible, into levels that align with the Common European Framework of Reference for languages (CEFR) to assess foreign language proficiency. The questions included in the test focused on four different tasks and five questions from each task were randomly selected for use in this experiment.

Task 1 tests knowledge of meaning and is designed to measure the test takers' ability to use phrase forms in order to understand the meanings from notices in a short text. QPT contained five questions for task 1, all of which were selected for this experiment. Task 2 tests knowledge of grammatical forms and is designed to measure the test takers' knowledge of grammar. In this task, test takers are asked to read a short gapped text and then complete the text by selecting one of three option choices. QPT contained five questions for task 2, all of which were also selected. Task 3 tests knowledge of pragmatic meaning and is designed to measure the test takers' knowledge of linguistic contextual information. It includes those verbal phenomena in which the gap between the literal meaning and the communicative meaning is clearly visible, and in which context plays a major role. Five questions out of the ten contained in QPT for task 3 were randomly selected for this experiment. Task 4 tests knowledge of form and meaning and is designed to test whether test takers can understand a long passage with gaps, as well as whether they have enough knowledge of grammar and vocabulary to correctly complete these gaps. Out of the twenty questions included in QPT for task 4, five were randomly selected. Questions were presented

to the participants ordered by their respective task, starting with five questions for task 1, five questions for task 2, five questions for task 3, and finally five questions for task 4.

No time restriction was given for answering each question. Upon answering, feedback was requested through the test's interface by asking the participants to characterise the previously answered question as "*Very Easy*", "*Easy*", "*Moderate*", "*Hard*", or "*Very Hard*". The experiment finished after answering all 20 questions and providing the respective feedback.

IV. Data Analysis

A. Participants

Twenty seven healthy individuals (20 male and 7 female), aged between 16 and 39 ($\mu_{\text{age}} = 27.3$, $\sigma_{\text{age}} = 5.8$), participated in this study by completing the English language test and providing their feedback while EEG, ECG, and EMG signals were recorded. Prerequisites for participating in the study were: (a) familiarity with basic use of a computer, and (b) basic understanding of the English language. Participants were recruited among international students from the University of the West of Scotland and from the local area (Paisley and Glasgow, Scotland, United Kingdom). The average duration of the experiment across all participants was $\mu_{\text{duration}} = 7.46$ min with a standard deviation $\sigma_{\text{duration}} = 1.89$ min and a maximum and minimum duration of 12.24 min and 3.72 min respectively.

B. Participants' test results and self-reported feedback

The test results and the self-assessed difficulty of the examined questions were analysed in order to evaluate the quality of the acquired data and to discover any visible trends. The English level of each participant was assigned according to the percentage of their correct answers, with <50% assigned to *Poor*, 50-60% to *Beginner*, 60-70% to *Elementary*, 70-80% to *Intermediate*, 80-90% to *Advanced*, and 90-100% to *Expert*. Following this convention, the majority of the 27 participants were assigned to levels between *Beginner* and *Advanced*, with only one participant assigned to *Poor* level. Interestingly, none of the participants were assigned to *Expert* level. The distribution of the assigned English levels for the 27 participants is depicted in the bar plot in Fig. 3. Furthermore the average percentage of correctly answered questions per question id is shown in Fig. 4 in the form of a scatter plot. From Fig. 4, it is evident that the questions included in the test used in this study demonstrate sufficient variation in the difficulty level, as observed by the distribution of successful answering across them.

The distribution (%) of the self-assessed difficulty for the questions answered by each participant in relation to the assigned English level is shown in Fig. 5. As expected, from Fig. 5, it is evident that for participants assigned from *Elementary*, to *Intermediate*, and then to *Advanced* levels, the percentage of questions self-assessed from *Very easy* to *Moderate* consistently increases (from 74.45% to 80% and then to 91.67% for each level respectively), while the percentage of questions self-assessed from *Hard* to *Very hard* consistently decreases (from 25.56% to 20% and then to 8.33%). However, this behaviour is not consistent for the *Poor*

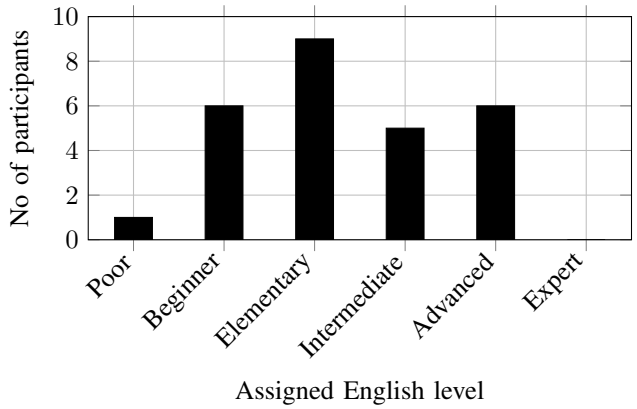


Fig. 3: Distribution of assigned English level.

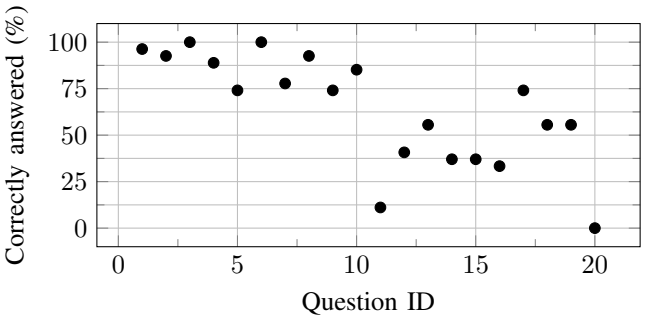


Fig. 4: Correctly answered questions (%) per question ID.

and *Beginner* levels. At the *Poor* level, the acquired data is not sufficient to extract reliable information since only one participant was assigned to that level. At the *Beginner* level, 65.84% of questions were self-assessed as *Very easy* or *Easy*. Considering that participants assigned to the *Beginner* level answered correctly less than 60% of the questions, it can be argued that they underestimated the difficulty of the asked questions due to their current English level.

Regarding the percentage of correctly answered questions in relation to the self-assessed difficulty levels, it is expected that the success rate will be lower the harder a question is considered. Indeed, as shown in Fig. 6a, the percentage of correctly answered questions decreases linearly (linear fit $R^2 = 0.986$) with the increase of the self-assessed difficulty level. The opposite trend can be noticed in the average time taken for the participants to answer each question in relation to the self-assessed difficulty level of the question. As shown in Fig. 6b, the average time spent for each question increases linearly (linear fit $R^2 = 0.968$) with the increase of the self-assessed difficulty level.

C. Physiological signals preprocessing

The acquired physiological signals were captured in a single continuous recording spanning the whole duration of the test for each participant. The timestamps associated with the signal samples and with the start and end of each question were then used in order to divide each recording into segments associated with one question each. Consequently, after removing the

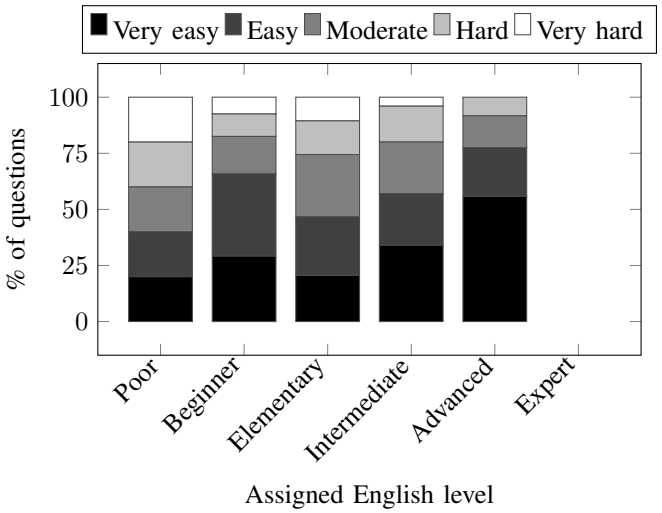
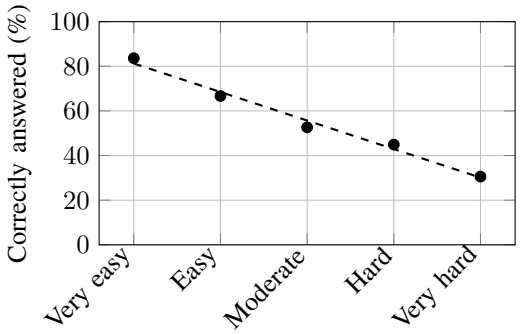
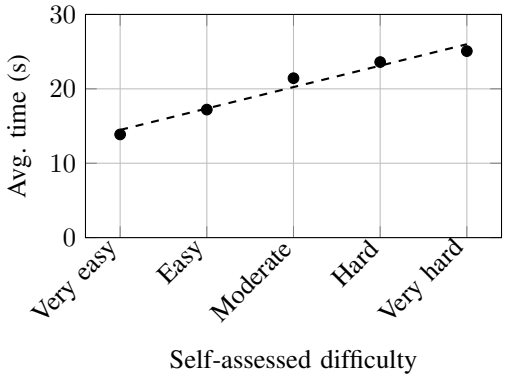


Fig. 5: Distribution (%) of self-assessed difficulty level of the questions in relation to the assigned English level. No participant was assigned to *Expert* level, while only one participant was assigned to *Poor* level.



(a)



(b)

Fig. 6: (a) Correctly answered questions (%) per self-assessed difficulty level. (b) Average time taken (s) to answer a question per question self-assessed difficulty level.

parts of the recording referring to before the start of the first question and after the end of the last question, each participant's recording was divided into twenty segments. This process led to the creation of 540 segments for each of the acquired physiological signals (27 participants $\times$ 20 questions). Furthermore, each segment was associated with the difficulty level assigned to each respective question by the participant, as well as with whether the respective participant answered it correctly.

Physiological signals are commonly contaminated with noise and artefacts, as a result of muscle movement, interference from electrostatic devices and power lines, cardiac activity, ocular artefacts (eye blinking, eye movement), etc. [59], [60]. To address this issue and reduce the effects of noise and artefacts in the performed analysis, the acquired signals were pre-processed before any further analysis. EMG signals were pre-processed as proposed in [61] by first cutting the peaks with values within the 3% of the lowest or highest values within the signal. Then, a 3rd order Butterworth FIR lowpass filter with a cutoff frequency of 0.4 Hz was applied and the resulting signal was normalised in the range [0,1].

To cope with the baseline wander that ECG signals suffer from due to movement or respiration, and with high-frequency noise caused by muscle activity [62], ECG signals were pre-processed by first removing baseline wander and then by filtering. For the removal of baseline wander, a median filter with a 200 ms window was first applied, followed by a median filter with a 600 ms window, and by subtracting the filtered signal from the original signal [63]. For further filtering, a bandpass filter between 0.7 - 20 Hz was applied to the already filtered ECG signal.

For the EEG signals, pre-processing was performed by first applying a Butterworth bandpass filter between 0.4 and 65 Hz. Then, the PREP EEG data pre-processing pipeline [64] was applied on the EEG signals using the EEGLAB toolbox [65]. The PREP pipeline consists of removing line-noise using filtering, referencing the EEG signal to an estimate of the "true" average reference, and finally the detection of "bad" channels and their replacement through interpolation in relation to the reference.

D. Feature extraction

After the pre-processing stage, the pre-processed segments of the recorded physiological signals were used in order to extract various statistical, spectral, and spatial features, to be used for the creation and evaluation of machine learning models. The following features were extracted form the EEG, ECG, and EMG signals:

1) Average PSD of EEG: The Power Spectral Density (PSD) of various frequency bands of EEG signals has been extensively utilised in EEG signal analysis, as it has been shown to correlate with human affective state [16]–[18], [41]. Following these works, the logarithm of the PSD of the low alpha (8-10 Hz), alpha (8-13 Hz), beta (13-30 Hz), gamma (30-64 Hz), and theta (4-8 Hz) frequency bands was computed for each of the 14 EEG channels. For the computation of the PSD of each channel, Welch's estimate of spectral power was first used, and then the FFT was computed across the component belonging to the analysed frequency band over a Hamming window of 2 s (512 samples) with 75% overlap (384 samples). The result

was then averaged in order to produce the PSD estimate. The final feature vector was created by concatenating the logarithm of the PSDs of each channel and frequency band, resulting to a vector with 70 features (14 channels $\times$ 5 frequency bands).

2) Band-based Spectral EEG features: Band-based spectral features are commonly used for biomedical signal analysis, having the advantage of low computational complexity once the spectrum (PSD) has been already computed. Using the previously computed PSD for the alpha, beta, gamma and theta bands of the EEG signal, the following five features were extracted from each band and channel of the EEG signal, as described by Monge-Álvarez *et al.* [66]: *Spectral Bandwidth*, *Spectral Crest Factor*, *Spectral Flatness*, *Spectral Roll-off*, and *Ratio f50 vs f90*. The final feature vector was created by concatenating the five spectral EEG features of each channel and frequency band, resulting to a vector with 280 features (14 channels $\times$ 4 frequency bands $\times$ 5 spectral features).

3) MFCC of EEG: Mel Frequency Cepstral Coefficients (MFCCs) have demonstrated promising results for EEG signal analysis [67]–[69], thus their performance was evaluated for the task at hand. Following Piciucco *et al.*'s [67] approach, 18 filterbanks were used for computing the MFCC features from each EEG channel, resulting to 12 cepstral coefficients per channel. The final feature vector was created by concatenating the cepstral coefficients of all channels, resulting to a vector with 168 features (14 channels $\times$ 12 cepstral coefficients). The MFCC features were computed for four different frequency bands of the EEG signal, namely 0.5-40 Hz, 4-40 Hz, 0.5-30 Hz, and 4-30 Hz.

4) Spatial and spectral ECG features: Previous research has shown that ECG-based features correlate with changes in human affective state [17], [18], [70], [71]. For example, happiness, fear, and sadness may lead to a decrease in heart rate variability (HRV) [72], while the peak heart rate may increase with pleasantness [73]. Following the approach in [17], 84 heart rate and HRV features were extracted from the ECG signals using the Augsburg Biosignal Toolbox (AuBT) [61] and were concatenated to create the final ECG feature vector. The extracted features were the amplitude's μ, $median$, σ, min, max, and $range$ (i.e. $max - min$) of the PQ, QS and ST complexes of the ECG signal and their first derivative, the number of intervals with latency > 50 ms from HRV, the PSD of HRV in the ranges $[0\,, 0.2]$ Hz, $[0.2\,, 0.4]$ Hz, $[0.4\,, 0.6]$ Hz and $[0.6, 0.8]$ Hz, and the μ, $median$, σ, min, max and $range$ of the HRV histogram.

5) Statistical EMG features: Based on previous research showing that affective states correlate with EMG signals [16], 21 statistical features were extracted from the EMG signals (x_{EMG}) using AuBT [61] and were concatenated to create the final EMG feature vector: μ, $median$, σ, min, max, and times per time unit that the signal reached the min and max, from x_{EMG}, x'_{EMG}, and x''_{EMG}.

6) Feature fusion: The fusion of features extracted from different physiological signal modalities has been shown to lead to increased performance in affect recognition studies [18], [44]. To this end, the previously described features were first normalised to the range $[0, 1]$ to compensate for their value range and various combinations of features were created by concatenating the respective feature vectors.

E. Classification experiments

Using the acquired data, four supervised classification experiments were designed. The first two attempted to use the features extracted from the acquired physiological signals in order to predict the self-assessed difficulty level of each answered question, by creating a separate classification model for each participant and by creating a global model that included all participants respectively. Both problems were converted to binary classification problems by grouping together samples assessed as *Very easy* and *Easy* for the first difficulty class (*Low*), as well as samples assessed as *Hard* and *Very hard* for the second difficulty class (*High*). Due to the number of difficulty levels being odd (5), the samples referring to the difficulty level in the middle (*Moderate*) were discarded since an even division was not possible. As a result, only 426 out of the 540 available samples ($\approx 79\%$) were used for these experiments. Converting multi-class classification problems to binary classification problems is common practice in the field of affective computing, as it usually results in improved classification accuracy (e.g. [16]–[18]).

The next two experiments attempted to use the features extracted from the acquired physiological signals in order to predict whether a participant would be successful in answering a question. Similar to the prediction of the questions' difficulty levels, one experiment focused on creating separate classification models for each participant, while the other on creating a global classification model using the data from all participants. Both problems are binary classification problems, with the class labels denoting whether a question was answered correctly or not (*True*/*False*). Furthermore, contrary to the prediction of the questions' difficulty levels, all 540 samples were used for the prediction of success in answering a question.

V. Experimental Results & Discussion

To the best of the authors' knowledge no work has attempted to establish the relation between physiological signals and the self-assessed difficulty level of answered test questions, as well as the success rate in answering the examined questions. Hence, a comparative study against other methods is not provided. Supervised classification experiments were conducted in order to distinguish between samples referring to (a) *Low* or *High* self-assessed difficulty, and (b) to samples referring to questions answered correctly or not (*True*/*False*). The examined classification algorithms were the k-Nearest Neighbour for $k = 1, 3, 5$, Linear SVM (LSVM), SVM with the Radial Basis Function kernel (SVM-RBF), Linear Discriminant Analysis (LDA), and Decision Trees (DT). The available implementations of MATLAB version R2018a were used for all the classification experiments. The classification performance of the trained models was evaluated in terms of Accuracy and F1-score. The F1-score considers both the Precision and Recall to compute the score, thus providing a superior classification performance metric than accuracy in

TABLE I: Best single-subject classification performance per feature for the prediction of self-assessed question difficulty.

Features	Classifier	Avg. Accuracy	Avg. F1-score	
ECG	LSVM	78.96	72.00	$\star\dagger\ddagger$
EMG	DT	76.80	69.99	$\star\dagger\ddagger$
EEG-PSDavg	1-NN	78.79	71.82	$\star\dagger\ddagger$
EEG-Spectral	LSVM	76.90	71.29	$\star\dagger\ddagger$
EEG-MFCC [4-40]	1-NN	77.50	71.26	$\star\dagger\ddagger$
EEG-MFCC [0.5-40]	LSVM	81.92	**74.21**	$\star\dagger\ddagger$
EEG-MFCC [4-30]	DT	78.02	71.89	$\star\dagger\ddagger$
EEG-MFCC [0.5-30]	LSVM	81.34	73.27	$\star\dagger\ddagger$
ALL	LSVM	81.20	72.68	$\star\dagger\ddagger$
ECG/EMG/EEG-PSDavg	1-NN	79.14	72.95	$\star\dagger\ddagger$
ECG-EMG	LSVM	79.65	71.99	$\star\dagger\ddagger$
EEG (ALL)	LSVM	82.48	74.10	$\star\dagger\ddagger$
n/a	Random	50.00	42.27	
n/a	Majority	80.75	44.20	
n/a	Class ratio	74.40	50.00	

$\star\dagger\ddagger$Statistically significant difference compared to random voting ($\star$), $p \leq 3.43 \cdot 10^{-4}$, majority voting ($\dagger$), $p \leq 1.34 \cdot 10^{-4}$, and voting according to the class ratio ($\ddagger$), $p \leq 0.011$.

TABLE II: Best single-subject classification performance per feature for the prediction of success in answering a question.

Features	Classifier	Avg. Accuracy	Avg. F1-score	
ECG	1-NN	59.26	51.49	$\dagger$
EMG	LDA	57.04	53.66	$\dagger$
EEG-PSDavg	DT	62.59	57.21	$\dagger$
EEG-Spectral	DT	62.04	56.46	$\dagger$
EEG-MFCC [4-40]	1NN	62.59	56.75	$\star\dagger\ddagger$
EEG-MFCC [0.5-40]	LSVM	63.52	56.33	$\dagger$
EEG-MFCC [4-30]	1-NN	59.81	52.55	$\dagger$
EEG-MFCC [0.5-30]	LSVM	66.67	**59.14**	$\star\dagger\ddagger$
ALL	LSVM	65.00	56.65	$\dagger$
ECG/EMG/EEG-PSDavg	LSVM	62.41	56.10	$\star\dagger$
ECG-EMG	DT	65.00	58.44	$\dagger$
EEG (ALL)	LSVM	65.93	57.89	$\dagger$
n/a	Random	50.00	48.13	
n/a	Majority	65.19	39.21	
n/a	Class ratio	56.89	50.00	

$\star\dagger\ddagger$Statistically significant difference compared to random voting ($\star$), $p \leq 0.0206$, majority voting ($\dagger$), $p \leq 0.0038$, and voting according to the class ratio ($\ddagger$), $p \leq 0.0153$.

cases of uneven class distribution. Furthermore, since the F1-score depends on which class is considered as positive, the reported F1-scores in this work are the average F1-scores between the examined classes. It must be noted that accuracy scores for the following experiments are reported only for reference purposes. Since the dataset is unbalanced (Difficulty: 75.35% *Low* vs 24.65% *High*, Success: 64.07% *True* vs 35.93% *False*) and the lack of additional samples does not permit further discarding of samples, the F1-scores from the following experiments provide a more reliable classification performance assessment that is not affected by class bias.

A. Single-subject classification

For the first set of experiments, separate models were trained for each subject in order to predict the self-assessed difficulty level of each answered question and in order to predict whether a participant would be successful in answering a question. A *Leave-One-Out* (LOO) cross validation procedure was applied in order to provide a fair performance evaluation of the trained models, avoid over-fitting, and compensate for the smaller number of samples available when only the samples for one participant are used. To this end, each subject-specific model was trained multiple times, each time tested with one sample and trained with the rest. After repeating this process for all samples of each subject, the average performance across all the iterations of the cross validation procedure was computed as the overall model's performance for each specific subject. Finally, the average classification performance across all subjects was reported. The previously described single-modality features, as well as feature fusion approaches, were evaluated and classification results in terms of average accuracy and average F1-score for the best performing settings are reported in Table I and Table II for the prediction of difficulty level and success in answering respectively.

For the single-subject models, the average classification F1-score for difficulty reached 74.21% using the EEG-based MFCC features for the 0.5-40 Hz band and the Linear SVM classifier. For the success in answering a question, the highest average classification F1-score (59.14%) was achieved using the EEG-based MFCC features for the 0.5-30 Hz band and the Linear SVM classifier. The use of feature fusion led to slightly lower average F1-scores, with the fusion of all the EEG-based features and the Linear SVM classifier achieving an average F1-score of 74.10% for predicting the difficulty level, and the fusion of ECG and EMG-based features with the Decision Tree classifier achieving an average F1-score of 58.44% for the prediction of the success in answering a question.

To test the acquired results for significance, they were compared to the analytically determined expected values for voting randomly (50% class probability), voting according to the ratio of classes (class probability equal to its ratio of samples within the set), and voting according to the majority class (100% probability of the majority class). It must be noted that the results for majority and class ratio voting are slightly overestimated since the class ratio would have to be estimated from the training set in each iteration of the cross-validation procedure [16]. The analytically computed results for the difficulty level and the success in answering are reported in Tables I and II respectively. A Wilcoxon signed-rank test was used to test for significance by comparing the distribution of F1-scores across each single-subject model for each setting reported in Tables I and II to the F1-scores distribution for random voting, majority voting, and voting according to the class ratio.

Random voting provided an expected average accuracy of 50% for both the difficulty level and the success in answering, and an average F1-score of 42.27% and 48.13% respectively. As can be seen in Tables I and II, the distribution of F1-scores was significantly different than random voting for all settings for the difficulty level prediction ($p < 3.43 \cdot 10^{-4}$), but only for three settings (EEG MFCC for 4-40 Hz, for 0.5-30 Hz, and the fusion of ECG, EMG, and EEG-PSDavg features) for the success in answering prediction ($p < 0.0206$). Results

TABLE III: Best multi-subject classification performance per feature for the prediction of self-assessed question difficulty.

Features	Classifier	Accuracy	F1-score
ECG	LSVM	76.53	**67.33** *†‡
EMG	LDA	74.18	56.41 *†‡
EEG-PSDavg	1-NN	67.14	55.76 *†‡
EEG-Spectral	DT	68.08	53.59 *†‡
EEG-MFCC [4-40]	DT	63.38	51.33 *†‡
EEG-MFCC [0.5-40]	LSVM	63.15	51.46 *†‡
EEG-MFCC [4-30]	3-NN	57.04	52.09 *†
EEG-MFCC [0.5-30]	LSVM	60.33	48.07 *†‡
ALL	LSVM	66.20	54.79 *†‡
ECG/EMG/EEG-PSDavg	LSVM	72.54	61.37 *†‡
ECG-EMG	LSVM	76.76	66.84 *†‡
EEG (ALL)	LSVM	62.91	54.17 *†‡
n/a	Random	50.00	46.57
n/a	Majority	75.35	42.97
n/a	Class ratio	62.85	50.00

*†‡Statistically significant difference compared to random voting ($\star$), $p \leq 0.0465$, majority voting ($\dagger$), $p \leq 5.77 \cdot 10^{-13}$, and voting according to the class ratio ($\ddagger$), $p \leq 1.56 \cdot 10^{-7}$.

TABLE IV: Best multi-subject classification performance per feature for the prediction of success in answering a question.

Features	Classifier	Accuracy	F1-score
ECG	5-NN	58.33	54.70 *†‡
EMG	1-NN	59.63	55.42 *†‡
EEG-PSDavg	DT	56.67	51.38 *†‡
EEG-Spectral	5-NN	55.56	49.43 *†‡
EEG-MFCC [4-40]	3-NN	56.85	51.79 *†‡
EEG-MFCC [0.5-40]	LSVM	57.22	52.70 *†‡
EEG-MFCC [4-30]	LSVM	55.93	51.08 *†‡
EEG-MFCC [0.5-30]	1-NN	58.15	**56.60** †‡
ALL	3-NN	57.78	54.35 *†‡
ECG/EMG/EEG-PSDavg	DT	57.22	54.28 *†‡
ECG-EMG	3-NN	59.07	***55.80*** *†‡
EEG (ALL)	LDA	54.63	53.48 †‡
n/a	Random	50.00	48.99
n/a	Majority	64.07	39.05
n/a	Class ratio	53.96	50.00

*†‡Statistically significant difference compared to random voting ($\star$), $p \leq 1.88 \cdot 10^{-4}$, majority voting ($\dagger$), $p \leq 4.13 \cdot 10^{-42}$, and voting according to the class ratio ($\ddagger$), $p \leq 3.06 \cdot 10^{-11}$.

for majority voting showed that the distribution of F1-scores was significantly different than majority voting for all settings for both difficulty level prediction ($p < 1.34 \cdot 10^{-4}$) and success in answering prediction ($p < 0.0038$). Furthermore, voting according to the class ratio provided an expected average F1-score of 50% for both the difficulty level and the success in answering, and an average accuracy of 74.40% and 56.89% respectively. Results showed that the distribution of F1-scores was significantly different than class ratio voting for all settings for the difficulty level prediction ($p < 0.011$), but only for two settings (EEG MFCC for 4-40 Hz and for 0.5-30 Hz) for the success in answering prediction ($p < 0.0153$).

B. Multi-subject classification

For the second set of experiments, all subject samples were used to create machine learning models for predicting the self-assessed difficulty level of each answered question and for predicting whether a subject would be successful in answering a question. To avoid over-fitting, remove any bias stemming from having samples from the same subjects in both training and test sets, and provide a fair comparison between the examined approaches, a *Leave-One-Subject-Out* (LOSO) cross validation procedure was applied. At each fold of the cross validation, all the samples related to a specific subject were used for testing the model and all the samples related to the other subjects were used for training. After repeating this process and testing the model for all the available subjects, the average performance across all iterations of the cross validation was computed as the overall performance of the model. Similar to the single-subject experiments, the previously described single-modality features, as well as feature fusion approaches, were evaluated. Classification results in terms of accuracy and F1-score for the best performing settings are reported in Table III and Table IV for the prediction of difficulty level and success in answering respectively.

For the multi-subject models, classification F1-score for the difficulty level reached 67.33% using the ECG-based features and the Linear SVM classifier, with the fusion of ECG and EMG-based features and the Linear SVM classifier providing a slightly lower F1-score of 66.84%. For the success in answering a question, the highest classification F1-score (56.60%) was achieved using the EEG-based MFCC features for the 0.5-30 Hz band and the 1-NN classifier, with the fusion of ECG and EMG-based features and the 3-NN classifier providing a slightly lower F1-score of 55.80%. For both tasks, the best performing single-modality features provided marginally better results than the best performing feature fusion approaches.

Similar to the single-subject experiments, the acquired results were tested for significance by comparing them to the analytically computed results for voting randomly, voting according to the ratio of classes, and voting according to the majority class. The analytically computed results for the difficulty level and the success in answering are reported in Tables III and IV respectively. To test for significance against random voting and class ratio based voting, unpaired Kruskal-Wallis tests were performed comparing the predicted class labels from random voting and class ratio based voting to the predicted labels for each experimental setting depicted in Tables III and IV respectively. To test for significance against majority voting, a paired Wilcoxon signed-rank test was used since the predicted class labels can be computed definitely on a one-by-one basis.

Random voting provided an expected average accuracy of 50% for both the difficulty level and the success in answering, and an average F1-score of 46.57% and 48.99% respectively. As can be seen in Tables III and IV, the performance of all settings was significantly different than random voting for the difficulty level prediction ($p \leq 0.0465$). For the success in answering prediction, all settings were significantly different than random voting ($p \leq 1.88 \cdot 10^{-}4$), apart from when the EEG MFCC for 0.5-30 Hz and the fusion of all EEG-based settings was used. Results for majority voting showed that the performance of all the settings was significantly different than

TABLE V: Signals/features that provided statistically significant results for the best performing classifier in each case, ranked from the highest to the lowest achieved F1-score.

Rank	Single-subject classification		Multi-subject classification	
	Difficulty	Success	Difficulty	Success
1	EEG-MFCC [0.5-40]	EEG-MFCC [0.5-30]	ECG	ECG-EMG
2	EEG (ALL)	EEG-MFCC [4-40]	ECG-EMG	EMG
3	EEG-MFCC [0.5-30]	-	ECG/EMG/EEG-PSDavg	ECG
4	ECG/EMG/EEG-PSDavg	-	EMG	ALL
5	ALL	-	EEG-PSDavg	ECG/EMG/EEG-PSDavg
6	ECG	-	ALL	EEG-MFCC [0.5-40]
7	ECG-EMG	-	EEG (ALL)	EEG-MFCC [4-40]
8	EEG-MFCC [4-30]	-	EEG-Spectral	EEG-PSDavg
9	EEG-PSDavg	-	EEG-MFCC [0.5-40]	EEG-MFCC [4-30]
10	EEG-Spectral	-	EEG-MFCC [4-40]	EEG-Spectral
11	EEG-MFCC [4-40]	-	EEG-MFCC [0.5-30]	-
12	EMG	-	-	-

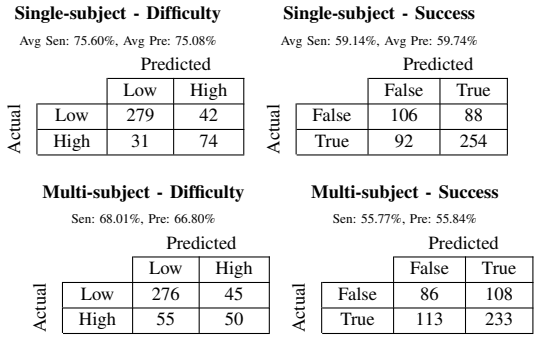


Fig. 7: Confusion matrices for the best performing settings

majority voting for both difficulty level prediction ($p \leq 5.77 \cdot 10^{-13}$) and success in answering prediction ($p \leq 4.13 \cdot 10^{-42}$). Furthermore, voting according to the class ratio provided an expected F1-score of 50% for both the difficulty level and the success in answering, and an accuracy of 62.85% and 53.96% respectively. Results showed that the performance of all settings was significantly different than class ratio voting for difficulty level prediction ($p \leq 1.56 \cdot 10^{-7}$), apart from when the EEG MFCC for 4-30 Hz features were used. For the prediction of the success in answering a question, all settings provided significantly different results than class ratio voting ($p \leq 3.06 \cdot 10^{-11}$).

C. Further discussion

Despite the non-significant results achieved for most best performing settings for the prediction of success in answering a question for the single-subject models, the overall best performing settings for both the prediction of difficulty level and success in answering provided statistically significant results, as shown in Tables I and II. Furthermore, from Tables III and IV, it is evident for the multi-subject models that while the best performing setting for the prediction of difficulty level (ECG-based features with the Linear SVM classifier) provided statistically significant results against all examined cases, the best performing setting for the prediction of success in answering (EEG MFCC for 0.5-30 Hz with the 1-NN classifier) failed the significance test against random voting. As a result, the second best performing setting (fusion of ECG and EMG-based features with the 3-NN classifier) must be considered as the actual best performing setting for the prediction of success in answering, since it passed all the significance tests. Consequently, the highest F1-score achieved for the prediction of success in answering a question for the multi-subject models was 55.80%, using the fusion of the ECG and EMG-based features and the 3-NN classifier. Confusion matrices (CFs), as well as average sensitivity and precision for the best performing setting for each examined problem are provided in Fig. 7. Similar to F1-scores, reported sensitivity and precision scores are the average sensitivity and precision scores between the two examined classes. It must be noted that for the single-subject approaches, CFs were created by aggregating the CFs of each single-subject model, thus metrics' values may deviate slightly when computed from the aggregated CF instead of the average metric across the different single-subject models.

Examining the results from Tables I, II, III and IV and from the ranking of the signals according to their performance in the examined problems in Table V, it is evident that EEG-based features provided the best performance for the single-subject approach. For the multi-subject approach, the best performance was achieved using ECG-based features for difficulty prediction and the fusion of ECG and EMG-based features for the prediction of success in answering a question. Interestingly, EEG-based features ranked last for the multi-subject approach, as seen in Table V. These findings indicate that EEG signals constitute a good descriptor for the examined tasks within a specific individual, having sufficient variation to allow the classifiers to differentiate between different affective states related to the examined problems. However, the variation across different individuals was insufficient to produce similar results across different subjects, where ECG and the fusion of ECG and EMG performed better. Nevertheless, a more in-depth study of the performance of such signals across different individuals would be required to extract safe conclusions.

Another additional point of concern is the variable length of the signal recordings used to extract the features. Unfortunately, due to different test takers taking a different amount of time to answer a question, it is not possible to determine the exact period of time that an affective response associated with the difficulty of a question or the success in answering occurred. As a result, the full duration of the signal recordings associated with each question was used, resulting into variable length data. Since some of the extracted features are dependent to the data length, and to establish whether the proposed methodology is affected by the use of variable length data, we repeated all the experiments conducted in this study for signals of the same duration. To this end, we first established the minimum answering time for the questions within the dataset, i.e. 3.2 s, and then we extracted the features described in Section IV-D from the last 3.2 s of each signal segment associated with a test question. Then, we repeated the four experiments described in Section IV-E. The achieved classification performance was comparable to the one achieved for the full duration of the recordings for each pair of features and classifier, indicating that the effect of variable length data was minimal. An example of the achieved F1-scores (%) for the 3.2 s segments, compared to the variable length segments, is shown in Table VI for the settings that provided the best performance when the full signal recordings were used.

TABLE VI: F1-scores (%) achieved using the 3.2 s signal segment size for the settings that provided the best performance using the full duration of the signal recordings.

Problem	Approach	Features	Classifier	F1-score Full	F1-score 3.2 s
Difficulty	Single-subject	EEG-MFCC [0.5-40]	LSVM	74.21	74.21
Difficulty	Multi-subject	ECG	LSVM	67.33	66.63
Success	Single-subject	EEG-MFCC [0.5-30]	LSVM	59.14	59.14
Success	Multi-subject	ECG-EMG	3-NN	55.80	55.35

Considering the overall results of this study, it is evident that single-subject models performed better than models containing multiple subjects. Nevertheless, the highest F1-score achieved for the multi-subject models, especially for the prediction of question difficulty (67.33%), allows the use of such models in practical applications. An ITS could potentially be equipped with a pre-trained multi-subject model that would be suitable for the general user. The pre-trained model could then evolve into a single-subject model through re-training with data gathered via user interaction. As a result, an ITS that follows this approach would not require training for each new user in order to exploit the affective state of the learners, thus being able to accommodate short-term users, while also being able to offer a more personalised experience to long-term users.

VI. Conclusion

In this work, we examined the potential use of EEG, ECG, and EMG physiological signals for affect detection during participation in a computerised English language test. Features extracted from recordings acquired from 27 individuals while answering twenty questions from the Oxford Quick Placement Test were used to train machine learning models for the task of predicting the self-reported difficulty level of each question and for predicting whether a question was answered correctly. Supervised classification experiments were conducted for both single-subject and multi-subject models using a multitude of features and classifiers.

For the single-subject models, the average classification F1-score for difficulty level prediction reached 74.21% using the EEG-based MFCC features for the 0.5-40 Hz frequency band and the Linear SVM classifier, while for the prediction of the success in answering a question, the average classification F1-score reached 59.14% using the EEG-based MFCC features for the 0.5-30 Hz frequency band and the Linear SVM classifier. For the multi-subject models, classification F1-score for difficulty level prediction reached 67.33% using the ECG-based features and the Linear SVM classifier, while for the prediction of the success in answering a question, classification F1-score reached 55.80% for the fusion of the ECG and EMG-based features and the 3-NN classifier. The statistical significance of the acquired results was tested against the random voting, majority voting, and class ratio voting classifiers resulting to statistically significant results for the reported F1-scores.

The acquired results provide evidence on the potential of physiological signals for the task of affect detection within the context of Intelligent Tutoring Systems (ITS). The success of both the single and multi-subject models, especially for the prediction of question difficulty, indicates that the proposed approach could be deployed within an ITS to assist in the personalisation and adaptation of the learning process according to the affective state of the learner, thus addressing to an extent the lack of a human tutor that could understand the affective state of the learners and adapt the learning process accordingly. The multi-subject model would be suitable as a generic model addressing all users, while single-subject models could be created and continuously evolved via the interaction with specific users. Furthermore, although it can be argued that the use of physiological signal sensors, like the ones used in this study, is intrusive and inconvenient for ITS users outside of a lab environment, the size and user-friendliness of such sensors is being continuously improved. Sensors are continuously becoming more wearable, more portable, more user-friendly, as well as cheaper. The proposed work attempted to provide a proof-of-concept that such bio-signal sensors could be successfully used in the context of ITS.

Future research will focus on examining the practicality and the performance of the proposed system within a real ITS environment. To this end, we plan to repeat the conducted experiments before and after providing training and tutoring to a group of students. The aim of that study will be two-fold; first to validate that our findings can be replicated when new data are used and secondly to examine whether the trained models for the examined students are stable across different recording sessions.

References

[1] X. Mao and Z. Li, "Agent based affective tutoring systems: A pilot study," *Comput. Edu.*, vol. 55, no. 1, pp. 202–208, 2010.
[2] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé, "When are tutorial dialogues more effective than reading?" *Cognit. Sci.*, vol. 31, no. 1, pp. 3–62, 2007.
[3] M. B. Ammar, M. Neji, A. M. Alimi, and G. Gouardères, "The affective tutoring system," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3013–3023, 2010.
[4] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Comput. Sci.*, vol. 104, pp. 437–444, 2017.
[5] C. N. Moridis and A. A. Economides, "Mood recognition during online self-assessment tests," *IEEE Trans. Learn. Technol.*, vol. 2, no. 1, pp. 50–61, Jan 2009.
[6] A. Landowska, "Affect-awareness framework for intelligent tutoring systems," in *HSI*, June 2013, pp. 540–547.
[7] H.-C. K. Lin, C.-H. Wu, and Y.-P. Hsueh, "The influence of using affective tutoring system in accounting remedial instruction on learning performance and usability," *Comput. Hum. Behav.*, vol. 41, pp. 514–522, 2014.
[8] N. Tsianos, Z. Lekkas, P. Germanakos, C. Mourlas, and G. Samaras, "An experimental assessment of the use of cognitive and affective factors in adaptive educational hypermedia," *IEEE Trans. Learn. Technol.*, vol. 2, no. 3, pp. 249–258, July 2009.
[9] K. Kiili and H. Ketamo, "Evaluating cognitive and affective outcomes of a digital game-based math test," *IEEE Trans. Learn. Technol.*, vol. 11, no. 2, pp. 255–263, April 2018.
[10] R. Picard, "Affective computing," MIT Media Laboratory Perceptual Computing Section, Tech. Rep. 321, 1995.
[11] R. W. Picard, "Affective computing: from laughter to IEEE," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 11–17, Sep. 2010.
[12] F. Alqahtani and N. Ramzan, "Comparison and efficacy of synergistic intelligent tutoring systems with human physiological response," *Sensors*, vol. 19, no. 3, p. 460, 2019.
[13] T. R. Lynch, A. L. Chapman, M. Z. Rosenthal, J. R. Kuo, and M. M. Linehan, "Mechanisms of change in dialectical behavior therapy: Theoretical and empirical observations," *J. Clin. Psychol.*, vol. 62, no. 4, pp. 459–480, 2006.

[14] M. I. Posner and M. K. Rothbart, *Educating the human brain*. American Psychological Association, 2007.
[15] C. Hattingh, J. Ipser, S. Tromp, S. Syal, C. Lochner, S. Brooks, and D. Stein, "Functional magnetic resonance imaging during emotion recognition in social anxiety disorder: an activation likelihood meta-analysis," *Front. Hum. Neurosci.*, vol. 6, p. 347, 2013.
[16] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, 2012.
[17] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health. Inf.*, vol. 22, no. 1, pp. 98–107, 2018.
[18] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, 2012.
[19] P. Lakhan, N. Banluesombatkul, V. Changniam, R. Dhithijaiyratn, P. Leelaarporn, E. Boonchieng, S. Hompoonsup, and T. Wilaiprasitporn, "Consumer grade brain sensing for emotion recognition," *IEEE Sens. J.*, vol. 19, no. 21, pp. 9896–9907, 2019.
[20] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sens. J.*, vol. 19, no. 6, pp. 2266–2274, 2019.
[21] J. A. Russell, "A circumplex model of affect." *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[22] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 1, no. e38, Feb. 2013.
[23] P. Sawangjai, S. Hompoonsup, P. Leelaarporn, S. Kongwudhikunakorn, and T. Wilaiprasitporn, "Consumer grade EEG measuring sensors as research tools: A review," *IEEE Sens. J.*, vol. 20, no. 8, pp. 3996–4024, 2020.
[24] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, "SHIMMER - A Wireless Sensor Platform for Noninvasive Biomedical Research," *IEEE Sens. J.*, vol. 10, pp. 1527–1534, Sept. 2010.
[25] J. A. Castro-García, A. J. Molina-Cantero, M. Merino-Monge, and I. M. Gómez-González, "An open-source hardware acquisition platform for physiological measurements," *IEEE Sens. J.*, vol. 19, no. 23, pp. 11 526–11 534, 2019.
[26] F. Alqahtani, S. Katsigiannis, and N. Ramzan, "ECG-based affective computing for difficulty level prediction in intelligent tutoring systems," in *UCET*, 2019, pp. 1–4.
[27] ——, "On the use of ECG and EMG signals for question difficulty level prediction in the context of intelligent tutoring systems," in *IEEE BIBE*, 2019, pp. 392–396.
[28] J. A. Kulik and J. D. Fletcher, "Effectiveness of intelligent tutoring systems: A meta-analytic review," *Rev. Educ. Res.*, vol. 86, no. 1, pp. 42–78, 2016.
[29] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," *J. Educ. Psychol.*, vol. 106, no. 4, pp. 901–918, 2014.
[30] B. D. Nye, "Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context," *Int. J. Artif. Intell. Educ.*, vol. 25, no. 2, pp. 177–203, Jun 2015.
[31] E. Mousavinasab, N. Zarifsanaiey, S. R. N. Kalhori, M. Rakhshan, L. Keikha, and M. G. Saeedi, "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods," *Interact. Learn. Envir.*, 2018.
[32] J. M. A. L. Andres, J. Ocumpaugh, R. S. Baker, S. Slater, L. Paquette, Y. Jiang, S. Karumbaiah, N. Bosch, A. Munshi, A. Moore, and G. Biswas, "Affect sequences and learning in betty's brain," in *LAK*, 2019, pp. 383–390.
[33] N. Bosch and S. D'Mello, "The affective experience of novice computer programmers," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 1, pp. 181–206, Mar 2017.
[34] R. Rajendran, S. Iyer, S. Murthy, C. Wilson, and J. Sheard, "A theory-driven approach to predict frustration in an its," *IEEE Trans. Learn. Technol.*, vol. 6, no. 4, pp. 378–388, Oct 2013.
[35] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion," in *IEEE ICALT*, Aug 2001, pp. 43–46.
[36] R. Zatarain-Cabada, M. L. Barrón-Estrada, J. L. O. Camacho, and C. A. Reyes-García, "Affective tutoring system for android mobiles," in *ICIC*, 2014.
[37] M. L. Barrón-Estrada, R. Zatarain-Cabada, J. A. Beltrán V., F. L. Cibrian R., and Y. H. Pérez, "An intelligent and affective tutoring system within a social network for learning mathematics," in *IBERAMIA*, 2012, pp. 651–661.
[38] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[39] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120 – 136, 2013.
[40] C. Marechal, D. Mikolajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, *Survey on AI-Based Multimodal Methods for Emotion Detection*. Cham: Springer, 2019, pp. 307–324.
[41] P. Arnau-González, M. Arevalillo-Herráez, and N. Ramzan, "Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals," *Neurocomputing*, vol. 244, pp. 81–89, 2017.
[42] A. Mert and A. Akan, "Emotion recognition from EEG signals by using multivariate empirical mode decomposition," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 81–89, Feb 2018.
[43] E. T. Pereira, H. M. Gomes, L. R. Veloso, and M. A. Mota, "Empirical evidence relating EEG signal duration to emotion classification performance," *IEEE Trans. Affective Comput.*, 2018, (Early Access).
[44] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affective Comput.*, vol. 6, no. 3, pp. 209–222, 2015.
[45] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, 2018, (Early Access).
[46] T. Althobaiti, S. Katsigiannis, D. West, and N. Ramzan, "Examining human-horse interaction by means of affect recognition via physiological signals," *IEEE Access*, vol. 7, pp. 77 857–77 867, 2019.
[47] W. Burleson, "Affective learning companions: strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance," Ph.D. dissertation, MIT, USA, 2006.
[48] P. Aghaei Pour, M. S. Hussain, O. AlZoubi, S. D'Mello, and R. A. Calvo, "The impact of system feedback on learners' affective and physiological states," in *ITS*, 2010, pp. 264–273.
[49] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D'Mello, "Affect detection from multichannel physiology during learning sessions with AutoTutor," in *AIED*, 2011, pp. 131–138.
[50] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 3, pp. 298–310, 2012.
[51] K. W. Brawner and B. S. Goldberg, "Real-time monitoring of ecg and gsr signals during computer-based training," in *ITS*, 2012, pp. 72–77.
[52] M. S. Hussain, H. Monkaresi, and R. A. Calvo, "Categorical vs. dimensional representations in multimodal affect detection during learning," in *ITS*, 2012, pp. 78–83.
[53] A. Dawood, S. Turner, and P. Perepa, "Affective computational model to extract natural affective states of students with asperger syndrome (AS) in computer-based learning environment," *IEEE Access*, vol. 6, pp. 67 026–67 034, 2018.
[54] G. A. Lujan-Moreno, R. Atkinson, and G. Runger, *EEG-based user performance prediction using random forest in a dynamic learning environment*. Nova Science Publishers, Inc., Jan. 2016, pp. 105–128.
[55] M. Chaouachi and C. Frasson, "Exploring the relationship between learner EEG mental engagement and affect," in *ITS*, 2010, pp. 291–293.
[56] E. Blanchard, P. Chalfoun, and C. Frasson, "Towards advanced learner modeling: Discussions on quasi real-time adaptation with physiological data," in *IEEE ICALT*, 2007, pp. 809–813.
[57] H. Ekanayake. (2015) P300 and Emotiv EPOC: Does Emotiv EPOC capture real EEG? Available at http://neurofeedback.visaduma.info/EmotivResearch.pdf.
[58] Oxford University Press, *Quick Placement Test*. Oxford University Press, 2001.
[59] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, 2004.
[60] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psy.*, vol. 25, no. 1, pp. 49–59, 1994.
[61] J. Wagner, "Augsburg biosignal toolbox (aubt)," *University of Augsburg*, 2005. [Online]. Available: https://www.informatik.uni-augsburg.de/en/chairs/hcm/projects/tools/aubt/

[62] M. Blanco-Velasco, B. Weng, and K. E. Barner, "ECG signal denoising and baseline wander correction based on the empirical mode decomposition," *Comput. Biol. Med.*, vol. 38, no. 1, pp. 1 – 13, 2008.
[63] N. Kannathal, U. R. Acharya, K. P. Joseph, L. C. Min, and J. S. Suri, "Analysis of electrocardiograms," in *Advances in Cardiac Signal Processing*. Springer, 2007, pp. 55–82.
[64] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K. Su, and K. Robbins, "The PREP pipeline: standardized preprocessing for large-scale EEG analysis," *Front. Neuroinform.*, vol. 9, p. 16, 2015.
[65] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.
[66] J. Monge-Alvarez, C. Hoyos-Barcelo, L. M. San Jose-Revuelta, and P. Casaseca-de-la-Higuera, "A machine hearing system for robust cough detection based on a high-level representation of band-specific audio features," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2319–2330, Aug 2019.
[67] E. Piciucco, E. Maiorana, O. Falzon, K. Camilleri, and P. Campisi, "Steady-state visual evoked potentials for EEG-based biometric identification," in *Proc. BIOSIG*, 2017, pp. 227–234.
[68] P. Nguyen, D. Tran, X. Huang, and D. Sharma, "A proposed feature extraction method for EEG-based person identification," in *ICAI*, 2012, pp. 826–831.
[69] E. Maiorana and P. Campisi, "Longitudinal evaluation of EEG-based biometric recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 5, pp. 1123–1138, May 2018.
[70] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *BI*, 2010, pp. 89–100.
[71] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *IEEE ICME*, 2005, pp. 940–943.
[72] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *Int. J. Psychophysiol.*, vol. 61, no. 1, pp. 5–18, Jul 2006.
[73] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, May 1993.