Predicting learners' effortful behaviour in adaptive assessment using multimodal data

Kshitij Sharma[1], Zacharoula Papamitsiou[1], Jennifer K. Olsen[2] and Michail Giannakos[1]
1. Norwegian University of Science and Technology, Trondheim, Norway
2. Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
kshitij.sharma@ntnu.no,zacharoula.papmitsiou@ntnu.no,jennifer.olsen@epfl.ch,michailg@ntnu.no

ABSTRACT

Many factors influence learners' performance on an activity beyond the knowledge required. Learners' on-task effort has been acknowledged for strongly relating to their educational outcomes, reflecting how actively they are engaged in that activity. However, effort is not directly observable. Multimodal data can provide additional insights into the learning processes and may allow for effort estimation. This paper presents an approach for the classification of effort in an adaptive assessment context. Specifically, the behaviour of 32 students was captured during an adaptive self-assessment activity, using logs and physiological data (i.e., eye-tracking, EEG, wristband and facial expressions). We applied k-means to the multimodal data to cluster students' behavioural patterns. Next, we predicted students' effort to complete the upcoming task, based on the discovered behavioural patterns using a combination of Hidden Markov Models (HMMs) and the Viterbi algorithm. We also compared the results with other state-of-the-art classification algorithms (SVM, Random Forest). Our findings provide evidence that HMMs can encode the relationship between effort and behaviour (captured by the multimodal data) in a more efficient way than the other methods. Foremost, a practical implication of the approach is that the derived HMMs also pinpoint the moments to provide preventive/prescriptive feedback to the learners in real-time, by building-upon the relationship between behavioural patterns and the effort the learners are putting in.

CCS CONCEPTS

• **Applied computing → E-learning**.

KEYWORDS

adaptive assessment, effort classification, multimodal learning analytics, hidden Markov models

ACM Reference Format:
Kshitij Sharma[1], Zacharoula Papamitsiou[1], Jennifer K. Olsen[2] and Michail Giannakos[1]. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Learners' on-task mental effort has been acknowledged for strongly relating to their educational outcomes, such as persistence in learning (e.g.,[5, 37]) academic achievement (e.g., [14, 55]), and learning performance (e.g., [42, 51]). Especially in adaptive learning and assessment contexts, in which the activities are tailored to fit learners' needs and mastery levels [12], learners' on-task mental effort is an important factor that reflects how actively they are engaged with the tasks and affects the correctness of activity outcomes and the quality of learning gains [64].

Learning (and assessment) adaptation has been found to have a positive impact on learners' engagement with activities (e.g., [5, 48]) and with the improvement of learners' performance (e.g., [42, 53]). In these settings, learners' active engagement with the adaptive tasks results in significant benefits, such as discouraging help-avoidance [47], improving problem-solving skills [34], increasing attendance and attention [21], improving self-regulation [31], guiding and facilitating autonomous learning decisions [53].

Active engagement is when the learners give their best efforts to understand and complete tasks - *effortful behaviour*, similar to *solution behaviour* proposed in [63]. Although active engagement leads to improved learning experiences [31, 42, 53], "true performance" can be potentially overshadowed and threatened by "hidden", *effortless behaviours*, commonly exhibited by learners. Effortless behaviour can be seen as a generalization of "rapid guessing"and "gaming the system" behaviours. The term *rapid guessing* was introduced to label when examinees rapidly respond to questions in a random fashion in Computerized Adaptive Testing (CAT) settings [63]. This kind of "cheating" behaviour is also apparent in adaptive learning contexts. For example, *"gaming the system"* behavioural patterns counterfeit the learning outcomes in intelligent tutors [6].

Timely detection and classification of "effortless" (e.g., [51, 72]) or "effortful" patterns (e.g., [5, 64]) is a prerequisite for preventing unwanted learner behaviour and improving adaptive learning service quality [8]. This paper suggests and evaluates an approach for classification of effort learners would exhibit in their next response, during an adaptive self-assessment process, using Hidden Markov Models and Viterbi algorithm with multimodal data.

2 RELATED WORK

2.1 Modeling the learning process

Modeling of the student learning process is not a new concept and is, in fact, a primary backbone to the research on intelligent tutoring and adaptive educational systems. Particularly, many of these models aim to predict the state of the students' knowledge to support these students in mastery learning.

One of the more widely used models in this space is Bayesian Knowledge Tracing (BKT) [17]. BKT uses a Hidden Markov Model (HMM) where the observable states are the student responses to questions (in terms of correctness) and the hidden states reflect the students binary knowledge of the skill. At each step, a student has a probability to transition from not knowing to knowing the skill. When they do not know the skill, there is a probability that they will guess correctly, and when they know the skill there is a probability that they will slip to make an error.

Since the original BKT model, many adaptations have been proposed to improve the prediction power of student correctness. Adaptations to BKT have allowed for student specific parameters [74] while others have incorporated student response time [41]. Rather than modeling each skill separately, Dynamic Bayesian Networks support the relationship between different skills [40]. Other models, such as Performance Factors Analysis, take a logistical regression approach to predict accuracy [57]. However, all of the these models focus primarily on predicting correctness.

On the other hand, other models have focused on predicting student behaviours during the learning process. When students are learning, the effort that they put into the process can be seen through certain behavioural patterns. For example, two main behaviours that are often tracked in adaptive systems are "wheel-spinning" [10] and "gaming the system" [7]. In "wheel-spinning" the students are putting in effort but are still not able to provide correct answers. This behaviour is important to differentiate between productive effort in which the process is still beneficial to the student [39]. By identifying "wheel-spinning" early, an intervention can be put into place. In contrast, when a student is "gaming the system" they are not providing any effort, yet the responses may look the same as a student who is "wheel-spinning". In this case, the intervention that is needed for the student would be different to match the effort of the student.

In addition to the cognitive behaviours, during the learning process, students also engage in self-regulatory behaviours. In technology-enhanced systems, students are often able to request hints. Models of this help-seeking behaviour can support the students' meta-cognitive skills [61]. More recently, the use of learning analytics to support meta-cognitive processes has been proposed [9]. For example, by combining learning curves and phases, groups with different self-regulated learning needs can be identified [46].

2.2 On-task mental effort in adaptive learning

Existing adaptation mechanisms take into consideration on-task learners' effort, as probabilities of effortful/less behaviours – from this point on, we will use the term "effortful behaviour" for both – i.e., to guess a solution or slip a correct answer [5, 8, 13, 30, 72]. Considering the required effort to successfully complete a task is a step ahead from taking into account only the task difficulty as effort reflects students' engagement with the task. In these approaches, a guessing parameter is incorporated in the learner models to describe the possibility of a learner responding correctly due to chance (effortless) instead of actively seeking to determine the correct answer (effortful) [30]. For instance, it was suggested that to identify features of an action using machine learning (a) an action should be characterized as a guess or slip immediately after it occurs and (b) information from subsequent actions should not be used [8]. An "effort-based" tutoring algorithm was proposed to inform adaptation decisions (i.e., selection of tasks and feedback) for each student on each task based on a student's number of incorrect answers, hints requested and response-time [5].

Most of the recent approaches to explain effortful behaviour usually rely on response time and guessing behaviour patterns detected in the log files of clickstream data from the learners' on-task activity [5, 13, 51, 52, 58, 70, 72]. It was found that repeatedly measuring mental effort (using subjective rating scales and associating the measurements with response-times) after performing individual tasks in a series was favoured for tasks that take longer than usual to complete [70]. Process mining techniques were also applied on response-time data to identify and model guessing behaviours [52].

Aforementioned contributions are based on the cognitive aspects of effort displayed by the students. Affective states, like boredom, have also been found to have a detrimental impact on learning outcomes [18]. Further, engaged concentration have been found to be positively associated with learning [56]. To detect learners' affective states while they interact with a given learning environment is necessary for adaptive learning technologies that aim to support and regulate learners' affect [24, 60].

2.3 Assessing student cognition and affective states through multimodal data

Recently, more sophisticated measurements have been employed for assessing a student's cognitive state. Multimodal data provide educational technology researchers with an unprecedented opportunity to gain insights and understanding of learners' actions in diverse learning contexts (e.g., [3, 23, 59, 64]). For instance, researchers found that electroencephalography (EEG) variables were sensitive to disengagement due to cognitive load [27]. Furthermore, effort-related cardiovascular responses can be mapped to success until a maximum effort has been achieved [73]. Speech, posture and gaze were used to automatically detect the moments when students' expectations are likely to influence their engagement [3]. Wristband data (e.g., Electrodermal Activity (EDA), Galvanic Skin Conductance (GSC), temperature) and accelerometer data were used to measure simultaneous arousal levels among students with respect to students' mood, motivation, affect and collaborative engagement [59], whereas the fusion of wristband data, gaze and emotions yielded highly accurate prediction of effort [64].

Previous studies show the use of multimodal data to estimate affect. Posture and interaction were used to detect affect, which was used to provide feedback to reduce students' frustration [20]. Learners' boredom, confusion, and frustration were detected by monitoring conversational cues, gross body language, and facial features [22]. An experiment comparing the affect-sensitive and non-affective tutors indicated that the affective tutor improved learning for low domain knowledge students, particularly at deeper levels of comprehension [22]. Three cognitive phases in problem solving, encoding, solving and responding, where found through an fMRI study in which an HMM was used to detect the stages from participants' brain activation patterns [67]. Another study provided EEG-based estimates of students' cognitive load and showed that EEG is a viable option to define the cognitive load of students [44].

3 METHODOLOGY

3.1 Research aim and question

The existing adaptation methods focus on estimating learners' effortful behaviour to complete tasks and use response-time indicators from clickstream data using probabilistic models [5, 8, 13, 30, 72]. However, as seen from the review of relevant literature [3, 23, 59, 64], to fully understand effort-related processes, learner on-task mental effort data need to be collected using multiple modalities.

To this end, this study goes a step beyond current state-of-the-art approaches and predicts the effort of the next response in adaptive assessment tasks, based on learners' previous behaviour, using machine learning and effort-related multimodal data produced during a learner's interaction with the adaptive system. This is important because it can contribute to improving the adaptation mechanism and to provide timely, proactive (cognitive, metacognitive or affective) feedback. This feedback can prevent, for example, students who are expected to give a wrong-effortless response to a task that is tailored to their ability or to encourage students who are predicted to give an effortful response to a task fitting their needs. It will use learners' previous behaviour in terms of attention, emotion, cognitive load, mental workload, load on memory, and arousal. Thus, the research question that guided this study was:

RQ: *"How can we predict learners' effort using multimodal data?"*

For answering the RQ and addressing the objective, this study suggests and evaluates the following approach: learners' states are captured using multimodal data (i.e., clickstreams, eye-tracking, wristband and EEG, and facial expressions, that have been acknowledged to satisfactorily explain effort-related behaviour) and their behaviour is coded by clustering those data. Learners' responses are then categorized in one of the two effort categories: effortful and effortless. Then, a combination of HMMs and Viterbi algorithm are used to predict the effort category of the learners' upcoming response based on their past behaviour, and the results are compared with other state-of-the-art classification methods.

3.2 Participants and Experimental Procedure

An online adaptive self-assessment activity was offered at a European University for a Web Technologies course (related to front-end development), and 32 undergraduate students (15 females [46.9%] and 17 males [53.1%], aged 18-21 years-old [M=19.24, SD=0.831]) were enrolled. The learners answered questions about based web-technology related to front-end application development (HTML, CSS). The questions were presented in a textual form via a simple GUI. The participants undertook the self-assessment activity individually, at an especially equipped and organized university lab for approximately 45 minutes.

Prior to their participation, all students signed an informed consent form that explained to them the data collection and the adaptive assessment procedure and gave the researchers the right to use the data collected for research purposes. After granting consent, the participants had to wear a wristband and an EEG cap and be connected to all the data collection devices (i.e., eye-tracker, wristband, EEG, cameras). Then, the actual adaptive self-assessment activity started and the students had to answer the tasks delivered to them one-by-one. Each task had two to four possible answers, but only one was correct. Every time the students submitted an answer to a task, their mastery class was revised and the next task was delivered according to the correctness of the answer and the discrimination ability of the tasks (briefly explained in the next sub-section). At the end of the procedure, the self-assessment score was available to the students, along with their full-test results, including all the tasks they had answered, their responses, the correctness of the responses, and the option to check the correct solution to the tasks for which they had submitted wrong answers, with a full explanation of the solution to support self-reflection. The experimental setup is illustrated in Figure 1.

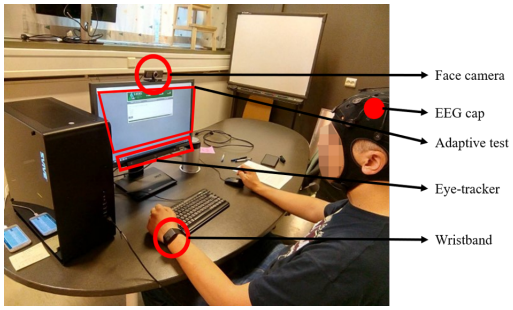

Figure 1: The experimental setup - The participant is connected to all data collection devices and is ready to take the self-assessment test

The participation to the procedure was optional. The adaptive self-assessment activity was offered to facilitate the students' self-preparation before the final exams, to help them track their progress, and self-reflect. The scores on the self-assessment had no influence to students' final grade in the course.

3.3 The adaptive self-assessment procedure

For adapting the self-assessment, Measurement Decision Theory (MDT) [62] was used to classify the students in three mastery classes based on their tasks responses (solutions), a priori task information, and a priori population classification proportions. The core of the methodology in use is the estimation of the students' mastery class every time they submit a solution. This estimation is reached by applying Bayes Theorem: $P(m_k|z) = c \cdot P(z|m_k) \cdot P(m_k)$, with: (a) $z = (z_1, z_2, ..., z_n)$: a student's response vector with $z_i \in \{0, 1\}$; (b) $P(m_k|z)$: the probability that the student belongs to mastery class m_k given z; (c) $P(z|m_k)$: the probability of responses z given the student's mastery class; (d) $P(m_k)$: the probability of a randomly selected student belonging to mastery class m_k; and (e) c: a standardization constant so that $P(m_1|z) + P(m_2|z) + P(m_3|z) = 1$.

At each step, the posterior classification probabilities $P(m_k|z)$ are treated as updated prior probabilities $P(m_k)$, and are used to help identify the next task to deliver. The selection of the next task is based on entropy, i.e., the selected task should maximise the reduction in entropy. This process continues until either a degree of decision accuracy is attained with a minimum number of tasks assigned, or a maximum number of tasks assigned is reached. In that case, for the termination of the test, if the Sequential Probability Ratio Test (SPRT) criterion [65] was not met after assigning 20

tasks, the self-assessment ended and the student was classified into the mastery class with the largest probability $P(m_k|z)$ to this point. Overall, a minimum of 10 and a maximum of 20 items were used to classify the students based on their diagnosed mastery level.

3.3.1 Preparation of the self-assessment tasks: The difficulty and discrimination ability of 80 multiple-choice tasks in total had been previously determined. The discrimination ability of a task corresponds to the probability of students in a given mastery class answering correctly to that task. For the tasks' discrimination ability configuration, three mastery classes of students were used (i.e., M1: final grade ≥ 7, M2: final grade ≥ 4, and M3: final grade < 4), and the respective probabilities were computed, using prior test results from 194 students who had already been classified, as follows: for each student and each task, we logged the correctness of each answer (right (1) - wrong (0), $P(z_1 = 0|m_k) = 1 - P(z_1 = 1|m_k)$). The probability $P(z|m_k)$ of the response vector z is the conditional probability of students in each mastery class responding correctly to each task, and equal to the product of the conditional probabilities of the task responses. $P(m_k)$ is an approximation of the portion of students in M1, M2 and M3. After the estimation of $P(m_k)$, and for each of the tasks, we estimated three probabilities, according to how likely a student in the given class m_k is to answer that task correctly. The tasks' difficulty levels were determined according to these probabilities by setting up the threshold values to 0.4 and 0.7 to discriminate between "easy", "medium" and "hard", respectively.

3.4 Data Collection

Students' self-assessment and interaction data were collected with a web-based self-assessment environment [50]. Furthermore, the following sensor data (multimodal data) were collected:

Eye-tracking: To record users' gaze we used the Tobii X3-120 eye-tracking device at 120 Hz sampling rate and using 5-point calibration. The device is non-invasive and mounted at the bottom of the screen. The screen resolution was 1920x1080 and the participants were 50–70 cm away from screen. Tobii's default algorithm was used to identify fixations and saccades (for details see [49]).

EEG: We recorded 20-channel EEG data organized in a standard 20 channel actiCAP layout following the international 10-20 system. We built upon previous studies that use EEG headsets in detecting cognitive engagement in the learning domain [32, 35, 66]. The raw EEG data was recorded at 500 Hz using a head-mounted portable EEG cap by ENOBIO (ENOBIO 20 EEG device), Fz was used as reference electrode, 2 channels were used for EOG correction, 1 channel for reference and 3 Channel Accelerometer sampling rate at 100 Hz. We applied an EOG filter to remove noise from blinks.

Face videos: Given the fact that we expected participants to exhibit minimal body and gesture information during the study, video recording focused on their face. We used a Logitech Web cam capturing video at 30 FPS. The webcam focus was zoomed 150% onto the faces of participants. The video resolution was 640x480.

Wristband: To record arousal data we used the Empatica E4 wristband. Participants wore the wristband on the non-dominant/non-playing hand. Four different measurements were recorded: 1) heart rate at 1 Hz, 2) electrodermal activity (EDA) at 64 Hz, 3) body temperature at 4 Hz, and 4) blood volume pulse at 4 Hz.

3.5 Measurements

For measuring students' behavioural states, the features in Table 1 were used. It should be noted that for EEG based features, first the features for each individual channel were computed, and then the average for all the 17 channels were computed as the actual features to be used. For pre-processing and data-synchronisation, we used the same steps mentioned in [64].

Attention: Eye-tracking data were used to compute attention. The most common practice to compute attention from students' gaze-patterns is to compute the average fixation duration during each sub-task [33, 69]. In this study, attention was computed using the average fixation duration over the time taken by the student to answer each self-assessment task.

Emotion: The facial data stream was used to compute and model students' emotional intensity. Extracting emotions from facial expressions is a common feature extraction technique [68]. First, the Facial Action Units (FAU) [25] were computed from the face videos, using the OpenFace Library [2]. The average of all FAU contributing to positive and negative emotions were calculated. Next, the average of the presence of high and low intensity emotions was computed. For example, happiness is a positive–low intensity emotion, while excitement is a positive–high energy emotion. Similarly, sadness is a negative–low intensity emotion, whereas anger is a negative–high intensity emotion. The absolute value of the emotional intensity was next calculated to capture the extent to which students externalized their emotions, regardless of their valance.

Cognitive load (CL): Decreasing alpha and theta band power[4] – to compute the cognitive load from the EEG signals, the following steps were followed. **a)** compute the discrete Fourier transform (DFT) of the signal; **b)** apply two band pass filters for computing the alpha (8–13 Hz) and theta (4–7 Hz) waves; **c)** extract the signal from the outputs of the filters by using an inverse DFT; **d)** compute the power per second for these two signals (power = root mean square of the amplitude); **e)** compute the mean of all the negative slopes for alpha waves and all the positive slopes of theta for the duration that the students take to respond to a given question.

Load on memory (LM): Theta band power[45] – use the steps **"a"** to **"d"** from the computation of cognitive load. For the theta waves take the mean band power computed for each second of the output.

Mental workload (MW): Alpha magnitude[11] – the average of alpha wave for the the duration that the students take to respond to a given question. To extract the alpha wave use the **"a"** to **"c"** from the computation of cognitive load, only for the alpha wave.

BVP: The mean blood volume pulse for the duration of a task.

HR: The mean heart rate for the duration of a task.

EDA: The mean electrodermal activation for the duration of a task.

TEMP: The mean skin temperature for the duration of a task.

3.6 Outcome Variable

Our outcome variable was the category of effort for a student's task-response. Effort is an indicator of how engaged the learners are in completing the tasks. In this study, the dichotomous index of tasks solution behaviour was adopted from [71] for measuring whether the students try to solve (effortful behaviour) instead of guessing the answer (effortless behaviour). The task was limited to multiple-choice questions, the learners did not spend a lot of time

Table 1: Definitions and sources for the computed features.

Measurement	Definition
Attention (ATT)	Average fixation duration [38]
Emotion intensity (EI)	based on Facial action Units
Cognitive load (CL)	Decreasing alpha and increasing theta band power [4]
Mental workload (MW)	Alpha magnitude [11]
Load on memory (LM)	theta band power [45]
Heart rate (HR)	Mean HR for a given question
Blood Volume Pulse (BVP)	Mean BVP for a given question
Electrodermal activation (EDA)	Mean EDA for a given question
Skin Temperature (TEMP)	Mean Temperature for a given question

on each question, the data collected was high frequency, but not nuanced enough to predict actual response times of the learners before they could answer (this would be a continuous version of classification problem). Therefore, we performed binary classification of effortful/effortless behaviour.

3.7 Data analysis

K-means was used to cluster students based on the multimodal data. The number of clusters was optimised using the within distance among the clusters. To characterise the clusters using the multimodal features, we used ANOVA to find out which features were the most distinguishing for each cluster. We conducted ANOVA with the cluster ID as the dependent and the features as the independent variable. Once the ANOVA yielded significant results, we further conducted pairwise one-way ANOVAs to check the affinity between the cluster ID and a specific feature.

Once the clusters and their characterising features were established, their IDs were used to label the observed states, and the effortful/effortless labels were used to refer to the hidden states of an HMM. The initial, transition and emission probabilities of HMM were initiated using uniform distributions. An expectation-maximisation algorithm was used to train the HMM with the first 10 responses of each student. The remaining responses were used for testing. The classification of effort of the next response was attained using the Viterbi algorithm. Viterbi is not a classification algorithm, but it is a generative algorithm. The input to Viterbi is an HMM and a sequence of N observations, and the algorithm generates the most probable sequence of N hidden states. To implement the classification method, a cluster ID sequence of length 11 was provided and the 11^{th} hidden state (from the most probable sequence of hidden states) was considered as the predicted effort category for the next response. In summary, K-means obtains the "observable" clusters from the multimodal data, the predicted classes are still "effortful/effortless" behaviour while solving the given problem.

HMM: this is a probabilistic model that is used to infer the hidden (unobserved data/states) from the observed states (usually data driven). The observed state are modelled based upon the Markov chains (with the assumption that the current observed state depends only on the previous observed state). Every HMM is characterised by three elements: the initial state, the state transition matrix (containing the probability of transition from one observed state to another observed state) and the emission matrix (containing the probability of inferring from one observed state, a hidden state). The transition and emission matrices could be obtained using a Expectation-maximisation algorithm (such as Baum-Welch or forward-backward algorithm). For mathematical details of the algorithm please refer to [15, 26].

Viterbi: the purpose of the Viterbi algorithm is to use a trained HMM and an observed sequence of states and generate the sequence of hidden states so that the joint probability of the sequence is maximised. The basic formula is as follows:

$$\mu(X_k) = \max_{X_{0:k-1}} P[X_{0:k}, Y_{0:k}] = \max_{X_{k-1}} \mu(X_{k-1}) P[X_k|X_{k-1}] P[Y_k|Y_{k-1}] \quad (1)$$

which results in a probability distribution for seeing different states given the observed data, with the objective to find the states that maximise the conditional probability of states given data.

4 RESULTS

4.1 Clustering

The number of clusters was selected using the within cluster distance. The left panel in Figure 2 plots the within distance vs. number of clusters. Using this plot, the optimal number of clusters is five, since using more than five clusters does not significantly decrease the within distance between the clusters. The right panel in Figure 2 shows the sample distribution in the clusters.

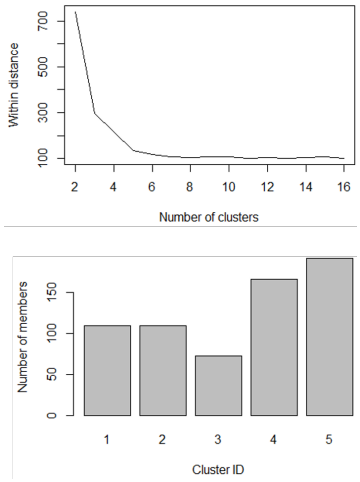

Figure 2: Results from clustering.

A pairwise ANOVA (Table 2) was conducted for each physiological feature (defined in Section 3.4) to guide the data-driven cluster definitions. The 4^{th} and 5^{th} columns of the table show an overall ANOVA (all five clusters). The 6^{th} to 15^{th} columns show the cluster ID with the significantly higher value for the feature in the first column. The missing values represent a non-significant difference. Let us take the example of "Attention". One can observe that there are two emerging patterns in this specific row: 1) all the pairwise

comparisons with cluster 2 show that cluster 2 has higher values in attention; and 2) all the pairwise comparisons with cluster 5 show that cluster 5 has lower values in attention. This implies that one of the defining features for cluster 2 is high attention, whereas one of the defining features of cluster 5 is low attention.

Using the same method for all rows in Table 2, the clusters are as follows. **C1**: high mental workload; high load on memory; low HR. **C2**: high attention; high cognitive load. **C3**: high EDA; high emotion; high HR; low mental workload; low ML. **C4**: low emotion; low cognitive load; high BVP. **C5**: low attention; high HR. Figure 3 illustrates the measures for the clusters.

4.2 Classification with Viterbi

Table 3 and Figure 4 show the resulting HMM. The most probable emission from a given state is presented in bold. Furthermore, the Viterbi classifications results are shown in Table 4. We notice a diagonal heavy confusion matrix for the test sample and hence high precision and recall for all the effort categories. The weighted means for precision and recall remain high (0.89 and 0.84, respectively).

4.3 Comparing Viterbi with SVM and RF

The classification results from Viterbi were compared against other classification methods. Table 5 shows four different implementations of machine learning classification approaches and the comparison to the proposed approach. A large comparative analysis [28] (179 general-purpose classification algorithms, 121 different datasets) found that RFs were the top-performing classification algorithm, only matched by kernel SVMs. This is the reason why we used only these two algorithms for comparison.

One way to predict the next effort category from the past data is to use the data from previous response as features, and the response on the next question as the target. The implementations in Table 5 use Support Vector Machines (SVM) with polynomial kernel and Random Forest classifiers, each of which have two different settings. In the first setting the previous response is not included as one of the features, while it is included in the second setting. The results show that only SVM with previous class information provides the closest results (although significantly lower classification quality) as the proposed method. Otherwise, all the other classifications yield significantly lower classification accuracy.

5 DISCUSSION

Adaptive learning and assessment systems aim to support learners by tailoring the delivered tasks to fit diagnosed learners' needs, skills, abilities and mastery levels [5, 8, 12, 58]. To achieve that, those systems consider learners' previous states, performance indices, task difficulty and an estimation of learners' on-task mental effort required to undertake and successfully complete the tasks. For the estimation of effort, probabilistic approaches and response-time patterns from clickstreams are commonly analyzed with a machine learning prediction algorithm [5, 8, 51], and the adaptation mechanism is updated accordingly. However, recent effort-related studies showcased that for a deeper and more holistic understanding and modelling of effort, clickstream data are not enough, and multiple other modalities should be considered [3, 21, 59, 64]. Therefore, the research question that guided this study was: *"How can we predict learners' effort using multimodal data?"*

To address our research question, this study suggested and evaluated a novel approach, using HMM and Viterbi algorithm with multimodal data for the classification of effort of learners' next response during an adaptive self-assessment activity and compared the classification results to other state-of-the-art methods. In this section, we elaborate on the findings from this study.

One might argue about the similarities between our approach and the most widely known method for adaptivity in education, that is, BKT. In our case, the HMM is implemented in a slightly different manner. As opposed to BKT, where the observable state is the correctness of the response, the observable states in our case are the physiological states characterised by the different MMLA measurements. Further, in BKT the hidden states are whether the student has the skill or not, while in our case the hidden states are whether the student has put in effort to solve the given problem or not. Furthermore, most of the work around effort uses the definitions of wheel-spinning [10] (effortful behaviour but incorrect answer) or gaming-the-system [7] (effortless behaviour and correct answer). We propose that effort is significant in itself to be considered as a standalone feature of the student learning process.

5.1 Explanation of the classification results and the HMM

The targeted outcome of this study was effort of next response. It is important to consider both the correctness of the response *and* the effort exhibited by the learners, because this combination of indices captures the *"true performance"*, i.e., how much the learners *truly try* to complete each task. In other words, effort captures on-task engagement, which is essential to understand the learning outcome, since they are correlated. The findings can be explained in terms of classification accuracy and the resulting HMM.

As seen in Table 4, the accuracy (weighted mean) achieved in the classification of effort of next response is high when employing effort-related multimodal data (e.g., eye-tracking, EEG, heart rate, EDA, BVP and facial expressions). This finding extends previous work on the estimation of on-task mental effort that employed response-time patterns from clickstreams or probabilities to guess/slip [5, 8, 30, 51], by efficiently fusing multimodal data.

Furthermore, since this study is the first one - to the best of our knowledge - to utilize mutlimodal data for the classification of effortful performance in adaptive assessment, we compared HMM-Viterbi with commonly used machine learning classification approaches to validate the appropriateness of the method. As seen in Table 5, since none of the other four methods (combination of the input data and the classification algorithm) could outperform the HMM-Viterbi combination, one can claim that the temporal modeling of the multimodal data (inherent in HMM-Viterbi) adds value to the classification of the effort categories.

The most intriguing finding of this study is the HMM itself. The underlying idea was to model learners' behaviour using clusters of effort-related multimodal data as the observed states of an HMM, and the effort of their next response as the hidden states of the HMM. Next, Viterbi was used to predict the next hidden state from the observed state. Table 3 illustrates the resulting HMM based on the method described above. The transition matrix is diagonal

Table 2: ANOVA results for the clusters. ATT = attention; EI = Emotional Intensity; CL = Cognitive Load; MW = Mental Workload; LM = Load on Memory; BVP = Blood Volume Pressure; EDA = Electrodermal Activation; TEMP = Skin Temperature.

			Overall ANOVA		Pairwise ANOVA between two clusters x-y the cell shows the cluster ID with larger values									
	Mean	SD	F-value	p-value	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
Attention	224.99	57.03	120.45	.00001	2	1	1	1	2	2	2	3	3	4
Emotional Intensity	0.40	0.08	26.77	.00001	2	3	-	5	3	2	2	3	3	5
Cognitive Load	1.72	0.15	50.16	.00001	2	-	1	-	2	2	2	3	-	5
Mental Workload	0.11	0.12	236.91	.00001	1	1	1	1	2	2	2	4	5	5
Load on memory	1.08	1.24	370.30	.00001	1	1	1	1	2	2	-	4	5	5
Heart Rate	80.22	10.45	164.75	.00001	2	3	4	5	3	2	5	3	-	5
Blood Volume Pressure	0.01	1.51	3.36	.009	-	-	-	-	-	4	-	-	-	4
Electrodermal Activation	0.11	0.17	390.87	.00001	2	3	-	5	3	-	2	3	3	4

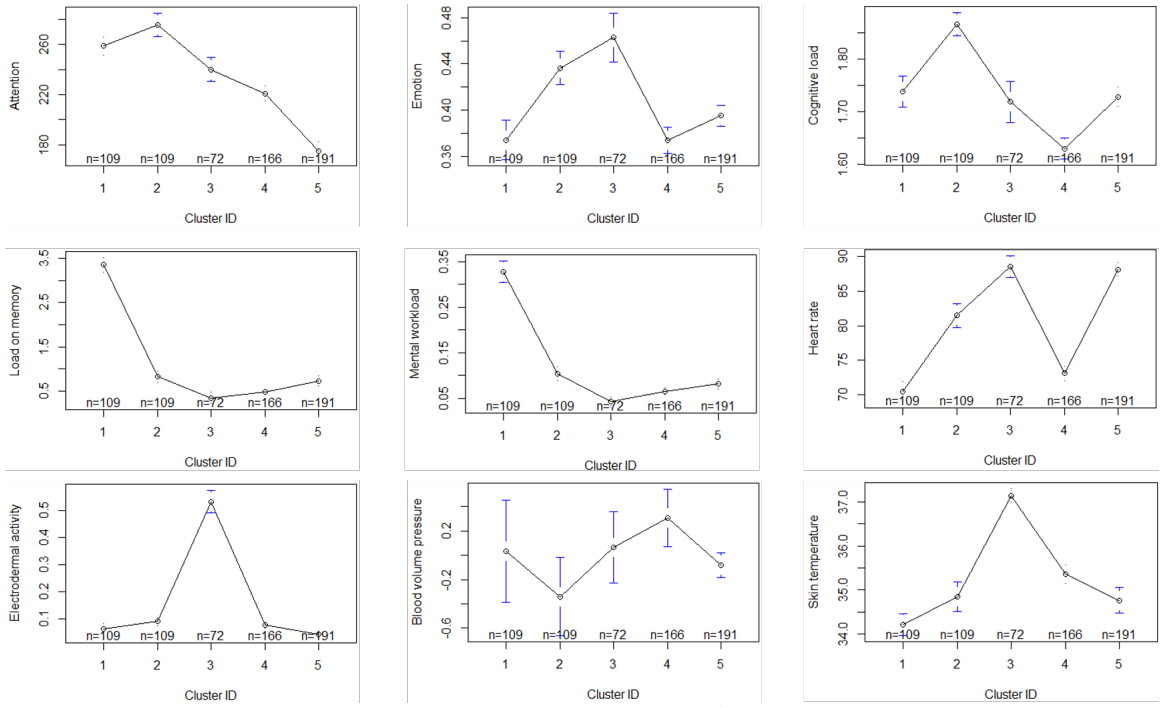

Figure 3: Results from ANOVA between clusters (n=number of members)

Table 3: HMM details

	Transition matrix					Emission matrix	
	C1	C2	C3	C4	C5	Guess	Solve
C1	0.81	0.11	0.01	0.02	0.03	**0.75**	0.25
C2	0.10	0.39	0.04	0.27	0.17	**0.72**	0.28
C3	0.01	0.05	0.83	0.01	0.07	0.29	**0.71**
C4	0.02	0.19	0.01	0.68	0.09	0.43	0.57
C5	0.02	0.09	0.01	0.09	0.77	**0.66**	0.34

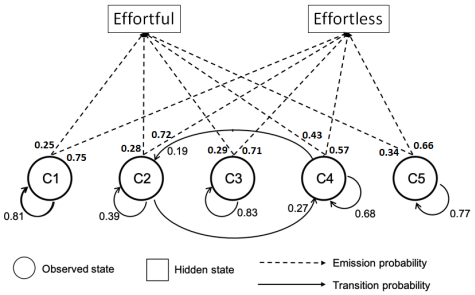

Figure 4: The resulting HMM, only the most probable emissions and the transitions more probable than 0.15 are shown.

heavy, i.e., the observed transitions are mostly making loops to the originating cluster, repeating themselves and indicating that the learners' physiological states (i.e., the clusters) do not change frequently. Given this facet and looking at the emission matrix, one can notice that each cluster is more strongly emitting to a single state than to multiple, i.e., the learners consistently exhibit a concrete type of effortful performance. For example, students

Table 4: Confusion matrix (test data) and the classification quality metrics.

	Predicted Class		classification Quality		
	Effortless	Effortful	precision	recall	f1-score
Effortless	107	5	0.92	0.82	0.93
Effortful	9	42	0.82	0.90	0.85
Wted. Mean			0.89	0.84	0.90

in C1 (high mental workload; high load on memory; low HR) or C5 (low attention; high HR) are more probable to respond with a effortless behaviour. This implies that those students might have been "trapped" into a behaviour that could hinder their success, and pinpoints the need for cognitive and/or affective feedback to push them out of this loop. In other words, one noteworthy observation from the emission matrix of the final HMM is that there is a clear affinity among the hidden and the observed states. Another interesting observation from the transition matrix of the final HMM is that C2 (high attention; high cognitive load) and C4 (low emotion; low cognitive load; high BVP) are the most probable to transit among each other. Furthermore, looking at the emission probabilities for C1, C2 and C5, one can observe that students in those two clusters are more likely to display *effortless* behaviour to the next task. This indicates that there is a significant probability that the students might end up in a loop of effortless attempts and signals the need for urgent proactive feedback to prevent such unwanted behaviour.

5.2 Implications for practice: opportunities for feedback

As mentioned earlier, there are moments to provide cognitive and/or affective feedback. Following are some suggestions for feedback.

Classified: Effortless behaviour – Observed: high mental workload & high load on memory & low HR (C1). When high mental workload and high load on memory are observed in the current state of the learner, it is likely that the predicted behaviour for the next (upcoming) task will be effortless. This is expected to happen because high mental workload and high load on memory drain the student's information processing capability, negatively affecting the motivation to perform the next task (the person experiences mental "exhaustness") [29]. In addition, the rate of cognitive processing becomes slower when low heart rates are observed [43]. This means that the learner who is experiencing low heart rates is possibly not able to sufficiently process the information of the upcoming task. In combination with the other observable physiological states, it is expected that this student will not exhibit high mental effort in that task, overall.

Classified: Effortless behaviour – Observed: high attention & high cognitive load (C2). Students in this state are focused on solving the tasks (high attention), however high cognitive load might lead to "mental fatigue" in certain situations [36], increasing the chances of displaying an effortless behaviour. In the moments when such behaviour is likely to happen, delivering (proactive) cognitive feedback might provoke learners to allocate more time to actively solve (effortful behaviour) the problem.

Classified: Effortful behaviour – Observed: high EDA & high emotion & high HR & low load on memory & low mental workload (C3). In this observed behaviour the combination of low mental workload and low load on memory indicate less chances of making mistakes [29]. In addition, both emotional intensity and EDA are high in this case. High emotional intensity suggests that the learner is experiencing strong emotions, e.g., excitement. At the same time, high EDA encompasses high emotional regulation [19], which means that in the observed state, the learners are capable of controlling their emotions to remain calm. When learners are in this observed state, they can remain "focused" on their efforts to solve the task. To keep them in the same physiological state, providing affective feedback praising the good work might work.

Classified: Non-confident – Observed: low emotion – low cognitive load & high BVP (C4). Students in this state are stressed (BVP is correlated to stress [1]) and might not be involved in effortful attempts, trying to avoid experiencing situations of high cognitive load. Stress and cognitive load might seem to be correlated, however their physiological measurements have been reported to be different [16]. It is not clear what might cause the effortless behaviour (e.g., if it is the stress they are experiencing or if they are not prepared/motivated enough to solve the tasks). On the other hand, the combination of low emotion (i.e., the learners have control of their emotions) and low cognitive load (i.e., the learners maintain their information processing capacity) might also result in an effortful behaviour. As such, it is also unclear what kind of feedback would be more effective in this case, and further work is required. This might be solved with a two step adaptive feedback in which the first step is to mitigate low high stress and the second step is to encourage low emotional intensity.

Classified: Effortless behaviour – Observed: low attention & high HR (C5). This state is straight forward: the learner is stressed (high HR indicates stressful situations [1]) and is not paying attention. In terms of providing proactive feedback, students should be prompted to pay more attention to questions and to remain calm.

6 CONCLUSIONS

This paper proposes and evaluates an approach for timely classification of learners' effortful/effortless behaviour during an adaptive assessment activity. Timely classification of such behaviour is one of the key requirements to prevent unwanted behaviour and improve learning gains[5, 8, 51, 64, 72]. To predict the effort categories in the adaptive assessment activity, we used a combination of HMM and Viterbi algorithm with the effort categories as the hidden states and the multimodal data-driven clusters as the observed states. The results show that the proposed method not only outperforms the contemporary classification algorithms but it also gives the educators several opportunities for providing (proactive) actionable feedback by pinpointing the exact moments in the learning activity where feedback is needed.

Effort is one of the factors influencing student performance among others possibilities. Our future work encompasses extending the proposed methods to other educational constructs, such as motivation and metacognition [54], and examining the generalizability of the method. Moreover, this paper provides the cues to provide feedback in terms of moments "when" to provide the feedback. "What" and "how" remain unanswered. To close the learning analytics loop effectively and efficiently, we aim to develop the feedback tools based upon our findings and further examine the effect of such a tool. Finally, many other multimodal features remain

Table 5: Comparing Viterbi with other classification methods

Method	Overall Precision	Overall Recall	Overall F1-score	F1 Comparison with HMM-viterbi for individual participants
HMM-viterbi	0.89	0.84	0.90	-
SVM polynomial with no previous class	0.80	0.77	0.79	$t(62) = 3.47$; $p < .05$
SVM polynomial with previous class	0.81	0.78	0.80	$t(62) = 3.04$; $p < .05$
Random forest with no previous class	0.76	0.68	0.73	$t(62) = 5.46$; $p < .05$
Random forest with previous class	0.79	0.73	0.76	$t(62) = 4.63$; $p < .05$

unexplored, which might also explain the engagement, motivation and metacognition related strategies of the learners with different levels of generalizability in terms of both the educational construct and learning activity.

To conclude, this paper opens the discussion towards merging multimodal physiological data for deeper understanding learners' effortful performance in adaptive settings, and detecting the moments for providing proactive cognitive and/or affective feedback accordingly to prevent learners from exhibiting unwanted and/or harmful behaviours.

REFERENCES

[1] Jonathan Aigrain, Michel Spodenkiewicz, Severine Dubuisson, Marcin Detyniecki, David Cohen, and Mohamed Chetouani. 2016. Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing* (2016).

[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6 (2016).

[3] Alejandro Andrade, Ginette Delandshere, and Joshua A. Danish. 2016. Using Multimodal Learning Analytics to Model Student Behaviour : A Systematic Analysis of Behavioural Framing. *Journal of Learning Analytics* 3, 2 (2016), 282–306.

[4] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review* 22, 4 (2010), 425–438.

[5] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. *Intl. Journal of Artificial Intelligence in Education* 24, 4 (2014), 387–426.

[6] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. Off-task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, 383–390.

[7] Ryan SJ Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[8] R S J d Baker, A T Corbett, and V Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Intelligent Tutoring Systems: 9th Intl. Conference, ITS 2008, Proceedings*, B P Woolf, E Aïmeur, R Nkambou, and S Lajoie (Eds.). Springer Berlin Heidelberg, 406–415.

[9] Maria Bannert, Inge Molenaar, Roger Azevedo, Sanna Järvelä, and Dragan Gašević. 2017. Relevance of learning analytics to measure and support students' learning in adaptive educational technologies. In *Proceedings of the Seventh Intl. Learning Analytics & Knowledge Conference*. ACM, 568–569.

[10] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *Intl. conference on artificial intelligence in education*. Springer, 431–440.

[11] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews* 44 (2014), 58–75.

[12] P Brusilovsky, S Somyürek, J Guerra, R Hosseini, V Zadorozhny, and P J Durlach. 2016. Open Social Student Modeling for Personalized Learning. *IEEE Transactions on Emerging Topics in Computing* 4, 3 (2016), 450–461.

[13] Shu-Ren Chang, Barbara S Plake, Gene A Kramer, and Shu-Mei Lien. 2011. Development and Application of Detection Indices for Measuring Guessing Behaviors and Test-Taking Effort in Computerized Adaptive Testing. *Educational and Psychological Measurement* 71, 3 (2011), 437–459.

[14] I-Shuo Chen. 2017. Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behavior* 72 (2017), 362–370.

[15] Kyoung Ho Choi and Jenq-Neng Hwang. 1999. Baum-welch hidden Markov model inversion for reliable audio-to-visual conversion. In *1999 IEEE Third Workshop on Multimedia Signal Processing (Cat. No. 99TH8451)*. IEEE, 175–180.

[16] Dan Conway, Ian Dick, Zhidong Li, Yang Wang, and Fang Chen. 2013. The effect of stress on cognitive load measurement. In *IFIP Conference on Human-Computer Interaction*. Springer, 659–666.

[17] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[18] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

[19] Hugo D Critchley. 2002. Electrodermal responses: what happens in the brain. *The Neuroscientist* 8, 2 (2002), 132–142.

[20] Jeanine A DeFalco, Jonathan P Rowe, Luc Paquette, Vasiliki Georgoulas-Sherry, Keith Brawner, Bradford W Mott, Ryan S Baker, and James C Lester. 2018. Detecting and addressing frustration in a serious game for military training. *Intl. Journal of Artificial Intelligence in Education* 28, 2 (2018), 152–193.

[21] Sidney D'Mello, Kristopher Kopp, Robert Earl Bixler, and Nigel Bosch. 2016. Attending to Attention: Detecting and Combating Mind Wandering During Computerized Reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1661–1669.

[22] Sidney D'Mello, Blair Lehman, Jeremiah Sullins, Rosaire Daigle, Rebekah Combs, Kimberly Vogt, Lydia Perkins, and Art Graesser. 2010. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *Intl. conference on intelligent tutoring systems*. Springer, 245–254.

[23] Sidney K D'Mello, Scotty D Craig, and Art C Graesser. 2009. Multimethod Assessment of Affective Experience and Expression During Deep Learning. *Int. J. Learn. Technol.* 4, 3/4 (2009), 165–187.

[24] Sidney K D'Mello, Amber Chauncey Strain, Andrew Olney, and Art Graesser. 2013. Affect, meta-affect, and affect regulation during complex learning. In *Intl. handbook of metacognition and learning technologies*. Springer, 669–681.

[25] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[26] David Elworthy. 1994. Does Baum-Welch re-estimation help taggers?. In *Proceedings of the fourth conference on Applied natural language processing*. Association for Computational Linguistics, 53–58.

[27] S H Fairclough, L J Moores, K C Ewing, and J Roberts. 2009. Measuring task engagement as an input to physiological computing. In *3rd Intl. Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–9.

[28] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[29] Daryl Fougnie and René Marois. 2007. Executive working memory load induces inattentional blindness. *Psychonomic bulletin & review* 14, 1 (2007), 142–147.

[30] Sujith M Gowda, Jonathan P Rowe, Ryan Shaun Joazeiro de Baker, Min Chi, and Kenneth R Koedinger. 2011. Improving Models of Slipping, Guessing, and Moment-By-Moment Learning with Estimates of Skill Difficulty. In *Educational Data MM*.

[31] Julio Guerra, Roya Hosseini, Sibel Somyurek, and Peter Brusilovsky. 2016. An Intelligent Interface for Learning Content: Combining an Open Learner Model and Social Comparison to Support Self-Regulated Learning and Engagement. In *Proceedings of the 21st Intl. Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 152–163.

[32] Mariam Hassib, Mohamed Khamis, Stefan Schneegass, Ali Sahami Shirazi, and Florian Alt. 2016. Investigating User Needs for Bio-sensing and Affective Wearables. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1415–1422.

[33] John M Henderson. 1992. Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition*. Springer, 260–283.

[34] Danial Hooshyar, Rodina Binti Ahmad, Moslem Yousefi, Moein Fathi, Shi-Jinn Horng, and Heuiseok Lim. 2016. Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education* 94 (2016), 18–36.

[35] Jin Huang, Chun Yu, Yuntao Wang, Yuhang Zhao, Siqi Liu, Chou Mo, Jie Liu, Lie Zhang, and Yuanchun Shi. 2014. FOCUS: enhancing children's engagement in reading by using contextual BCI training sessions. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1905–1908.

[36] Susanne M Jaeggi, Martin Buschkuehl, Alex Etienne, Christoph Ozdoba, Walter J Perrig, and Arto C Nirkko. 2007. On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience* 7, 2 (2007), 75–89.

[37] Yeonji Jung and Jeongmin Lee. 2018. Learning Engagement and Persistence in Massive Open Online Courses (MOOCS). *Computers & Education* 122 (2018), 9–22.

[38] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[39] Shimin Kai, Nicole Shechtman, Ma Victoria Almeda, Cristina Heffernan, Ryan S Baker, and Neil Heffernan. 2017. Modeling wheel-spinning and productive persistence in Skill Builders. In *Workshop and Tutorials Chairs*. ERIC, 5.

[40] Tanja Käser, Severin Klingler, Alexander G Schwing, and Markus Gross. 2017. Dynamic Bayesian networks for student modeling. *IEEE Transactions on Learning Technologies* 10, 4 (2017), 450–462.

[41] Chen Lin, Shitian Shen, and Min Chi. 2016. Incorporating student response time and tutor instructional interventions into student modeling. In *Proceedings of the 2016 Conference on user modeling adaptation and personalization*. ACM, 157–161.

[42] Min Liu, Emily McKelroy, Stephanie B Corliss, and Jamison Carrigan. 2017. Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development* 65, 6 (2017), 1605–1625.

[43] 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.

[44] Caitlin Mills, Igor Fridman, Walid Soussou, Disha Waghray, Andrew M Olney, and Sidney K D'Mello. 2017. Put your thinking cap on: detecting cognitive load using EEG during learning. In *Proceedings of the seventh Intl. learning analytics & knowledge conference*. ACM, 80–89.

[45] P Missonnier, M-P Deiber, G Gold, P Millet, M Gex-Fabry Pun, L Fazio-Costa, P Giannakopoulos, and V Ibáñez. 2006. Frontal theta event-related synchronization: comparison of directed attention and working memory load effects. *Journal of Neural Transmission* 113, 10 (2006), 1477–1486.

[46] Inge Molenaar, Anne Horvers, and Ryan S Baker. 2019. Towards hybrid human-system regulation: Understanding children'SRL support needs in blended classrooms. In *Proceedings of the 9th Intl. Conference on Learning Analytics & Knowledge*. ACM, 471–480.

[47] Mary Muir and Cristina Conati. 2012. An Analysis of Attention to Student – Adaptive Hints in an Educational Game. In *Intelligent Tutoring Systems*, Stefano A Cerri, William J Clancey, Giorgos Papadourakis, and Kitty Panourgia (Eds.). Springer Berlin Heidelberg, 112–122.

[48] Nur Baiti Afini Normadhi, Liyana Shuib, Hairul Nizam Md Nasir, Andrew Bimba, Norisma Idris, and Vimala Balakrishnan. 2019. Identification of personal traits in adaptive learning environment: Systematic literature review. *Computers & Education* 130 (2019), 168–190.

[49] A Olsen. 2012. The Tobii I-VT fixation filter: Algorithm description [White paper]. *Retrieved from Tobii Technology from http://www. tobii-ipro. com/siteassets/tobiipro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vtfixation-filter. pdf* 2012 (2012).

[50] Z.K. Papamitsiou and A.A. Economides. 2013. Towards the alignment of computer-based assessment outcome with learning goals: The LAERS architecture. In *2013 IEEE Conference on e-Learning, e-Management and e-Services, IC3e 2013*.

[51] Z. Papamitsiou and A.A. Economides. 2015. A temporal estimation of students' on-task mental effort and its effect on students' performance during computer based testing. In *Proceedings of 2015 Intl. Conference on Interactive Collaborative Learning, ICL 2015*.

[52] Z. Papamitsiou and A.A. Economides. 2016. *Process mining of interactions during computer-based testing for detecting and modelling guessing behavior*. Vol. 9753. Springer, Cham.

[53] Zacharoula Papamitsiou and Anastasios A Economides. 2019. Exploring autonomous learning capacity from a self-regulated learning perspective using learning analytics. *British Journal of Educational Technology* (2019).

[54] Zacharoula Papamitsiou, Anastasios A Economides, and Michail N Giannakos. 2019. Fostering LearnersâĂŹ Performance with On-demand Metacognitive Feedback. In *European Conference on Technology Enhanced Learning*. Springer, 423–435.

[55] A Pardo, F Han, and R A Ellis. 2017. Combining University Student Self-Regulated Learning Indicators and Engagement with Online Learning Events to Predict Academic Performance. *IEEE Transactions on Learning Technologies* 10, 1 (2017), 82–92.

[56] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2014. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics* 1, 1 (2014), 107–128.

[57] Philip I Pavlik, Hao Cen, and Kenneth R Koedinger. 2009. Performance Factors Analysis–A New Alternative to Knowledge Tracing. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*. IOS Press, 531–538.

[58] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education* 98 (2016), 169–179.

[59] H J Pijeira-Díaz, H Drachsler, P A Kirschner, and S Järvelä. 2018. Profiling sympathetic arousal in a physics course: How active are students? *Journal of Computer Assisted Learning* 34, 4 (2018), 397–408.

[60] Jennifer Robison, Scott McQuiggan, and James Lester. 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *2009 3rd Intl. conference on affective computing and intelligent interaction and workshops*. IEEE, 1–6.

[61] Ido Roll, Ryan S Baker, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. 2005. Modeling students' metacognitive errors in two intelligent tutoring systems. In *Intl. Conference on User Modeling*. Springer, 367–376.

[62] L M Rudner. 2003. The classification accuracy of Measurement Decision Theory. In *Annual meeting of the National Council on Measurement in Education*. Chicago.

[63] Deborah L Schnipke and David J Scrams. 2002. Exploring issues of examinee behavior: Insights gained from response-time analyses. In *Computer-based testing: Building the foundation for future assessments*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 237–266.

[64] Kshitij Sharma, Zacharoula Papamitsiou, and Michail N. Giannakos. 2019. Building Pipelines for Educational Data: Using AI and Multimodal Analytics to Explain Learning in Adaptive Self-Assessment. *British Journal of Educational Technology* (2019).

[65] Judith A Spray and Mark D Reckase. 1996. Comparison of SPRT and Sequential Bayes Procedures for Classifying Examinees Into Two Categories Using a Computerized Test. *Journal of Educational and Behavioral Statistics* 21, 4 (1996), 405–414.

[66] Daniel Szafir and Bilge Mutlu. 2013. ARTFul: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1001–1010.

[67] Caitlin Tenison, Jon M Fincham, and John R Anderson. 2016. Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive psychology* 87 (2016), 1–28.

[68] Vasileios Terzis, Christos N. Moridis, and Anastasios A. Economides. 2013. Measuring Instant Emotions Based on Facial Expressions During Computer-based Assessment. *Pers. Ubiquit. Comput.* 17 (2013), 43–52.

[69] Pieter JA Unema, Sebastian Pannasch, Markus Joos, and Boris M Velichkovsky. 2005. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual cognition* 12, 3 (2005), 473–494.

[70] Tamara van Gog, Femke Kirschner, Liesbeth Kester, and Fred Paas. 2012. Timing and Frequency of Mental Effort Measurement: Evidence in Favour of Repeated Measures. *Applied Cognitive Psychology* 26, 6 (2012), 833–839.

[71] Steven L Wise and Xiaojing Kong. 2005. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education* 18, 2 (2005), 163–183.

[72] Steven L Wise, Megan R Kuhfeld, and James Soland. 2019. The Effects of Effort Monitoring With Proctor Notification on Test-Taking Engagement, Test Performance, and Validity. *Applied Measurement in Education* 32, 2 (2019), 183–192.

[73] Rex A Wright and Leslie D Kirby. 2001. Effort determination of cardiovascular response: An integrative analysis with applications in social psychology. In *Advances in Experimental Social Psychology*. Advances in Experimental Social Psychology, Vol. 33. Academic Press, 255–307.

[74] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, H Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik (Eds.), Vol. 7926 LNAI. Springer Berlin Heidelberg, Berlin, 171–180.