Investigating Multi-Modal Measures for Cognitive Load Detection in E-Learning

Nico Herbig
nico.herbig@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus

Tim Düwel
tim.duewel@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus

Mossad Helali
mossad.helali@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus

Lea Eckhart
Patrick Schuck
first.last@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus

Subhabrata Choudhury
subha@robots.ox.ac.uk
University of Oxford

Antonio Krüger
antonio.krueger@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus

ABSTRACT

In this paper, we analyze a wide range of physiological, behavioral, performance, and subjective measures to estimate cognitive load (CL) during e-learning. To the best of our knowledge, the analyzed sensor measures comprise the most diverse set of features from a variety of modalities that have to date been investigated in the e-learning domain. Our focus lies on predicting the subjectively reported CL and difficulty as well as intrinsic content difficulty based on the explored features. A study with 21 participants, who learned through videos and quizzes in a Moodle environment, shows that classifying intrinsic content difficulty works better for quizzes than for videos, where participants actively solve problems instead of passively consuming videos. Regression analysis for predicting the subjectively reported level of CL and difficulty also works with very low error within content topics. Among the explored feature modalities, eye-based features yield the best results, followed by heart-based and then skin-based measures. Furthermore, combining multiple modalities results in better performance compared to using a single modality. The presented results can guide researchers and developers of cognition-aware e-learning environments by suggesting modalities and features that work particularly well for estimating difficulty and CL.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *User studies*; *Empirical studies in ubiquitous and mobile computing*; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

cognitive load detection; e-learning; multi-modality; wearables

ACM Reference Format:
Nico Herbig, Tim Düwel, Mossad Helali, Lea Eckhart, Patrick Schuck, Subhabrata Choudhury, and Antonio Krüger. 2020. Investigating Multi-Modal Measures for Cognitive Load Detection in E-Learning. In *28th ACM Conference on User Modeling, Adaptation and Personalization(UMAP'20), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340631.3394861

UMAP '20, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-6861-2/20/07…$15.00
https://doi.org/10.1145/3340631.3394861

1 INTRODUCTION

The e-learning industry is continuously growing, with a predicted compound annual growth rate of 7% until 2025 [34]. While modern e-learning systems offer a variety of customization possibilities, they neglect the user's current cognitive load (CL), which can strongly influence the content or speed that is appropriate for his/her current state (see e.g. [62]). Here, we see CL as "a variable that attempts to quantify the extent of demands placed by a task on the mental resources we have at our disposal" [8]. In contrast to e-learning platforms, human teachers in traditional learning try to estimate their student's CL and react to it by asking follow-up questions or adding additional explanations. Similarly, cognition-aware e-learning systems could provide further clarifying contents when a high CL is detected, or decide to move on to more complex topics when the CL drops. Furthermore, informing the instructor of an online course about the learners' cognitive states could help improve the learning material and tailor it to individual needs.

These and other adaptations aiming to keep the learner in the optimal range of CL [54] would be possible if e-learning systems had the ability to estimate the cognitive state of a user. Plenty of approaches to measure cognitive load, stress, etc. relying on one or few sensors have been proposed in the literature and allow some form of cognition awareness [4, 45, 50]. Often the sensors used in these works are nowadays even integrated into consumer devices like smartwatches, making the concepts feasible in practice. However, the interplay between those individual sensors and the power of using multiple modalities simultaneously remain underexplored.

In this paper, we investigate the so far most diverse set of cognitive load measures in the e-learning domain, including heart, skin, eye, body posture, performance, and subjective measures. Furthermore, a study using a realistic e-learning setting, where participants learn through videos and quizzes, is presented. Based on the captured data, we analyze how well predictive models using feature combinations from the explored modalities can predict intrinsic difficulty as well as the perceived CL and difficulty. In particular, we analyze which sensor modalities (eye, heart, skin) are more or less suitable for estimating CL, thereby guiding researchers and developers of future cognition-aware e-learning systems.

2 RELATED WORK

This section discusses related studies by giving an overview of CL measures and presenting literature on cognition-aware (e-)learning.

2.1 Overview of Cognitive Load Measures

Cognitive load theory [43, 56] has been developed in psychology and is concerned with an efficient use of people's limited cognitive resources to apply acquired knowledge and skills to new situations [42]. CL theory distinguishes intrinsic CL (i.e., the difficulty of the task, like a simple arithmetic addition compared to solving an integral equation), germane CL (i.e., the construction of learning schemas), and extraneous CL (i.e., load introduced by a bad design of learning materials) [42]. However, usually the total amount of CL is measured because accurately distinguishing the three remains an unsolved problem [33]. Approaches to measure CL can be roughly divided into four categories: subjective measures, performance measures, behavioral measures, and physiological measures.

Subjective measures are based on the assumption that subjects can self-assess and report their cognitive processes after performing a task [43]. Several scales exist, and introspection is often used as ground truth to evaluate how well CL can be assessed by other means, such as physiological measurements.

Performance measures assume that when working memory capacity is overloaded, a performance drop occurs due to the increase in overall CL [8]. However, by increasing their efforts, humans can compensate for the overload and maintain their performance over a period of time at the cost of additional strain and fatigue [21].

Behavioral measures can be extracted from user activity while performing a task. In the context of e-learning, this could for example be mouse and keyboard input [1] or the head pose [2].

Last, a lot of research has been done on *physiological measurements*, which assume that human cognitive processes can be observed in human physiology [30]. Eye-tracking is frequently used for physiological CL measurements: the pupil diameter increases with higher CL [25, 41], the frequency of rapid dilations changes [11], and the blink behavior adapts [58]. Furthermore, [10] as well as [55] showed that fixations and saccades can also be used for CL predictions. Apart from the eyes, the skin also provides information about the user's cognitive state: galvanic skin response (GSR) can be used to determine whether a user feels stressed [61] and provides information about the CL [51]. Remote measurements of the skin temperature have also been effective [28, 66], where the nose temperature drops upon high workloads. Further commonly used indicators rely on the cardiovascular system: blood pressure [66], heart rate [39], and especially heart rate variability (HRV) [46] have been shown to correlate with CL. Other physiological measures include respiration [7] and brain activity [23, 52].

Multi-modal approaches, investigating a variety of sensors simultaneously, have also been presented: Guhe et al. [16] combine heart, skin, eye, and behavioral measures to estimate the workload during an N-back task. In the translation domain, Vieira [60] analyzes how eye-based, typing-based, time-based, and subjective measures relate to each other in a multivariate analysis, while Herbig et al. [19] extend this feature set to also include heart and skin measures and use it to investigate how well these can predict subjective CL ratings.

2.2 Cognition-Aware (E-)Learning

Several forms of adaptive (e-)learning, often also called intelligent tutoring systems, have been proposed in the literature: Kuo et al. [31] propose the idea of a context-aware learning system that considers factors like facial expressions, human voice, or body temperature. Recommendations of learning content based on ontologies about the learner and the content, as well as behavioral, positional, temporal, and technological data, have also been proposed [44, 67]. Furthermore, dynamic user interface adaptations [14] and adaptive visualizations [9], driven by physiological parameters, were suggested to support learning. The concept of affective e-learning, which uses emotion feedback to improve the learning experience, was proposed in [49]. The work showed in a feasibility study that biosensors can be utilized for this purpose. A review of affective computing in education can be found in [64], which highlights the essential role that positive emotion has on comprehension performance. Bahreini et al. [4] investigate emotion recognition using webcams and microphones to better respond to the affective states of students, as human teachers would in traditional learning. Similarly, Ishimaru et al. [26] link eye tracking data, including fixations and pupil diameter as well as thermography, to surveys about cognitive states when studying a digital physics book. Based on this, they propose to provide individualized information to enhance learning abilities. Leony et al. [32] show that such adaptations can affect cognitive processes like memorization and decision-making. A framework for learning analytics based on wearable devices "to capture learner's physical actions and accordingly infer learning context" is proposed in [35]. They implement student engagement detection for the classroom based on arm movement to intervene when engagement is low, or to provide incentives when it is high. Moissa et al. [37] review the literature on measuring students' effort and propose that students could also be alerted when they should take a break or move to less challenging tasks as detected through wearables. Finally, participants of an online survey [20] have misgivings regarding the use of sensor data in e-learning unless opt-in mechanisms, etc., are well-considered. Furthermore, the participants suggested a variety of potential adaptions of e-learning platforms towards the learner's CL, e.g., adapting the content's speed or level of detail, or splitting it into parts of different lengths.

Apart from these works focusing mainly on conceptual design or correlations, several works go a step further and investigate the feasibility of adaptations to CL by training predictive models: For arithmetic calculations, Borys et al. [6] train trinary classification models (low CL, high CL, without task) based on brain and eye

activity. Similarly, a Ridge regression model based on brain activity during arithmetic operations determines the task complexity with comparatively low error [54]. This model was then used in a second study to adaptively propose tasks to learn arithmetic additions in the octal number system [63]. Similarly, Galán and Beal [13] use SVMs to predict the success or the failure of students solving math problems based on a combination of attention and workload signals from EEG sensors. Instead of EEG data, Mock et al. [36] use touchscreen interactions to classify CL for children solving math problems.

We further extend upon these previous works by (i) incorporating even more sensors and features and analyzing their impact on predictive models, (ii) exploring a more realistic e-learning setting, where students actually watch videos and solve quizzes instead of artificially performing one mental calculation after another, and (iii) particularly focusing on the differences in performance between different modalities (eye, heart, skin), thereby guiding researchers and developers of future cognition-aware e-learning systems.

3 USER STUDY & MEASURES OF CL

To test which measuring approaches can actually reflect different levels of CL in e-learning, we capture data from a variety of sensor modalities a user study[1].

3.1 Procedure, Apparatus, and Content Used

3.1.1 Overview. Participants first fill in a data protection form and a pre-questionnaire. Then, they go through six pairs of mathematics videos and corresponding quizzes in counter-balanced order within a Moodle e-learning platform. After each quiz, there is a small break task, and at the very end, a final questionnaire is filled out.

3.1.2 Pre-questionnaire. The initial questionnaire captures demographics, previous e-learning experience, and information about last night's sleep, as well as perceived exhaustion and tiredness. Last, the math background of the participants is captured to (a) confirm that they match our targeted group (see Section 4.1), and (b) to see if effects found might depend on differences in prior knowledge.

3.1.3 Apparatus. Then the learner is equipped with a Microsoft Band v2 on her right wrist, a Garmin Forerunner 935 sports watch and an Empatica E4 wearable on the left wrist (the Garmin is further up), a Polar H7 heart belt on her chest, and a Tobii eye tracker 4C with Pro SDK, as well as a web-cam and a Microsoft Kinect v2 camera facing her. As input possibilities, keyboard and mouse are used, and a 22-inch monitor displays the Moodle environment.

3.1.4 Videos. We chose 3 mathematical topics for the experiment: vectors, integration, and eigenvectors. For each topic two videos are presented, one considered easy as it is part of the curriculum for the high school certificate, and one considered hard as it is part of the university's "mathematics for computer scientists" curriculum. Note here that this distinction into easy/hard is based solely on the concept of intrinsic CL, while extraneous and germane load also depend on how the material is taught [56]. We aimed to make the teaching style as comparable as possible by using videos from the popular German math Youtube channel "Mathe by Daniel Jung", thereby ensuring the videos to have the same speaker, presenting in a very similar fashion, being filmed from the same perspective, etc. The length of the videos is roughly 5 minutes each (mean=312s, min=247s, max=368s). After each video, participants quickly assess their CL as well as the content difficulty (see Section 3.2.1).

3.1.5 Quizzes. Afterward, and similar to [26], participants take a multiple choice quiz of 2 to 4 questions on the previously watched content, testing whether they understood what they saw. In contrast to the videos, which are consumed rather passively, the quizzes ensure that participants actively work. The questions were created by the authors, then refined after discussions with two students matching our participant profile, and afterward tested in a pre-study with two participants. While we cannot guarantee that the quizzes are didactically 100% comparable, the question design and the iterative testing aimed to make the quizzes as consistent as possible. Participants also rate each quiz on the same subjective scales as the videos (see Section 3.2.1).

3.1.6 Break Task. Following each quiz, participants engage in a break task to limit the interference between content items. As a task, participants connect numbers drawn on paper in increasing order [59] while verbally stating each number, to clear both the visual as well as the verbal working memory (see Baddeley's model of working memory [3]).

3.1.7 Post-questionnaire. At the very end, participants fill out a final questionnaire, which again captures tiredness, stress, and exhaustion, as well as motivation, to be compared to these factors before the experiment, thereby analyzing tiredness effects. Furthermore, participants judge the relative differences in difficulty between the three content topics.

3.2 Analyzed Measures of Cognitive Load

This section gives a brief overview of the vast amount of measures of CL analyzed in the study.

3.2.1 Subjective Measures. Within the e-learning platform, we ask participants for two ratings after each piece of content: for estimating subjective CL (SubjCL), the commonly used scale proposed by Paas and Van Merriënboer [43] is utilized. The single 9-point question is 'In solving or studying the preceding problem I invested' with a choice of answers ranging from 'very, very low mental effort' to 'very, very high mental effort'. Furthermore, the difficulty measure proposed by Kalyuga et al. [27], which is a 7-point scale, ranging from 1 (extremely easy) to 7 (extremely difficult), asking about the difficulty of the task, is analyzed (SubjDiff).

3.2.2 Performance Measures. While the time required to watch a video is not relevant, due to the constant duration of the video, we analyze the quiz time, where we expect more difficult quizzes to require more time. Furthermore, the percentage of quiz questions answered correctly is used as a measure of performance.

3.2.3 Behavioral Measures. As a behavioral measure, the body posture is captured by a Microsoft Kinect v2. Based on the skeleton, we calculate the distance from the head to the screen, hypothesizing that learners come closer for harder content.

[1] The study was approved by Saarland university's ethical review board.

3.2.4 Physiological Measures. As physiological measurements, we integrate eye-, heart-, and skin-based measures in our experiment.

For *eye-based features*, the web-cam, which is naturally not as precise as the eye tracker but easily accessible on most modern devices, is used to calculate the eye aspect ratio, which indicates the openness of the lids [53]. The remote Tobii 4C eye tracker with Pro SDK records the raw gaze positions. Based on this raw data, we calculate the amount of blinking (of less than 2s length) and the number of fixations and also normalize this by the content time [58]. We further compute the fixation durations and saccade durations [12, 38], all of which have been shown to be indicators of CL. Furthermore, we calculate the probability of visual search proposed in [15]. Last, the eye tracker also captures the pupil diameter [41] which we use to compute higher-level features: first blinks from the signal are replaced by linear interpolation; then, the Index of Cognitive Activity, which is the frequency of small rapid dilations of the pupil [11] that was shown to be more robust to changes in illumination, is calculated based on this signal. Two approaches are implemented: one uses a wavelet transformation to calculate the number of rapid dilations, while the other simply counts how often a sample deviates by more than 5 times the rolling standard deviation from the rolling mean of the signal. Last, we also implemented the work of [22], which checks for sharp changes and continuations of the ramp in the Hilbert unwrapped phase of the pupil diameter signal.

For *heart measures*, we capture the heart rate from both the Polar belt and the Garmin watch. The Polar belt, as well as the Empatica wristband, further capture the RR interval, which is the length between two successive Rs (the peaks) in the ECG signal. Based on this, we calculate the often-used CL measures of heart rate variability (HRV) [46], in particular, the root mean square of successive RR interval differences and the standard deviation of NN intervals. The NN intervals normalize across the RR intervals and thereby smooth abnormal values. Furthermore, we add the additional HRV features NN50 and pNN50, which are the number and percentage of successive NN intervals that differ in duration by more than 50 ms [48], for both the Empatica wristband and the Polar belt, to the analysis. The Empatica wristband also measures the blood volume pulse (BVP), which is the change in volume of blood measured over time. The BVP amplitude [24], which contains the amplitude between the lowest (diastolic point) and highest (systolic point) peak in a one-second interval, as well as the median absolute deviation and the mean absolute difference among the BVP values [17], are used as features.

For *skin-based features*, the Microsoft Band and Empatica wristbands both measure the galvanic skin response, which is an indicator of CL. We also transform this signal to the frequency domain as described in [8]. In accord with their work, we calculate data frames of length 16, 32, and 64 samples, which are similarly transformed to the frequency domain and normalized by the participant average. Furthermore, we use the Ledalab software[2] to calculate higher-level skin conductance features on the Empatica raw data. It provides us with "global" features, namely the mean value and the maximum positive deflection, and "through-to-peak (TTP)/min-max" analysis, namely the number of significant (i.e., above-threshold) skin conductance responses (SCRs), the sum of SCR amplitudes of significant SCRs, and the response latency of the first significant SCR. Furthermore, we use Ledalab to perform a Continuous Decomposition Analysis (CDA) [5], which separates skin conductance data into continuous signals of tonic (background) and phasic (rapid) activity. The features based on this CDA analysis again include the number of significant SCRs, the SCR amplitudes of significant SCRs, and the latency of the first SCR. Furthermore, the average phasic driver, the area of the phasic driver, and the maximum phasic activity as well as the mean tonic activity features are created by the Ledalab software. Finally, the Empatica and Garmin devices also measure skin temperature, which we use as another feature.

3.2.5 Data Normalization and Content-Wise Feature Calculation. The features described above can be categorized into two classes: *single features* and *continuous features*.

(1) *Single features* yield only one value per content item: this class comprises subjective measures, time measures, quiz performance measures, the amount-based eye features, and all Ledalab skin features. However, one should note that the time and performance features here really can only be calculated on the whole content, while the amount-based eye and skin features could also be calculated over shorter periods of time.

(2) All other features are a *continuous signal* (of different sampling rates) that we transform into a directly usable set of values per content. Each signal is first normalized as described in [8] by dividing it by the participant's mean value. Then 5 features are calculated from this normalized signal: the average, standard deviation, minimum, maximum, and range (max − min), which is comparable to many related works, e.g., [6] and [26].

Given all our single features and calculating these 5 subfeatures for the continuous ones, we have 202 features values per video/quiz content per participant. We manually inspected the data distribution per content item and participant for outliers and overall data quality. Values were filtered according to visual inspection and related literature: data above 100000 kΩ for the raw skin resistance, as well as Polar RMSSD above 300, SDNN values above 250 [57], and finally HR and RR samples which fall outside the acceptable 50–120 beats per minute or 500–1200 ms ranges were removed [48].

4 DATA ANALYSIS RESULTS & DISCUSSION

We first analyze the questionnaires and look at the subjective ratings and time required for individual content items, thereby validating that the chosen method for data acquisition can be used as planned.

Then we analyze how well our three main metrics, (1) subjective CL (SubjCL, regression in range 1 to 9), (2) subjective difficulty (SubjDiff, regression in range 1 to 7), and (3) intrinsic difficulty (IntrinDiff, binary classification), can be estimated based on the captured sensor data.

The last part of the analysis aims to better understand which feature modalities consistently perform better or worse than others, thereby providing suggestions on how to implement cognition-aware e-learning systems in practice. For this, we present results from multiple analyses, including intraclass correlation coefficients and an analysis of how the performance of the predictive models changes when we leave out different modalities.

[2] http://www.ledalab.de/, accessed 24/01/2020

4.1 Participants and Questionnaire Results

Overall 21 students, aged 20 to 33 years (mean=25.2), participated (m=17). Roughly half (9 participants) described their e-learning experience as rather good or good and used platforms like Moodle, Udacity, or Coursera. To ensure a comparable background, we required all participants to be enrolled in a computer-science-related course of study, and to have already successfully passed the mathematics lectures covering our selected topics. Furthermore, participants had to self-assess their background in the three chosen topics on 5-point scales, where they claimed to have the most prior knowledge for the topic of vectors (mean=3.38, SD=0.81), closely followed by integration (mean=3.14, SD=0.85), and last, eigenvectors (mean=2.67, SD=0.97). In the post-questionnaire at the very end, participants were asked to rate the three topics in terms of difficulty on a 7-point scale: vectors were rated the easiest (mean=2.19, SD=1.12), followed by integration (mean=3, SD=1.18), and eigenvectors (mean=3.05, SD=1.28), where 3 corresponds to "rather easy". According to a univariate ANOVA for the three topics, vectors are significantly easier than the other two topics, which are on the same level ($F(2,40) = 4.84$, $p < .05$). This means that we should also investigate each topic separately and not only analyze the differences between all easy and all hard content items. Using an ANCOVA to test if these differences only come from a higher prior knowledge in the topic of vectors shows that this is not the case ($F(2,18) = 0.37$, $p = .693$ for interaction between topic and prior knowledge).

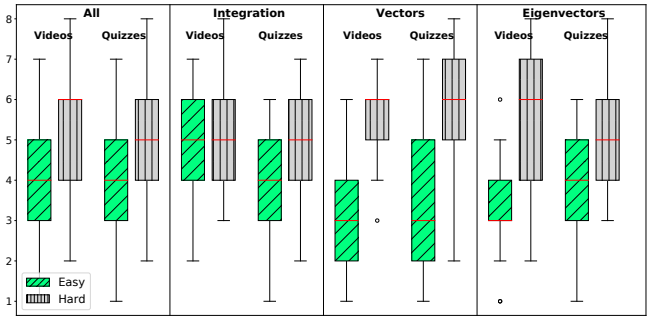
The current tiredness (mean=2.57, σ=0.98), exhaustion (mean=2.05, σ=0.87), and stress (mean=2.0, σ=0.95, all ratings on a 5-pt scale) were in an acceptable state at the beginning of the experiment. The corresponding values after the experiment (tiredness (2.58/5, 0.96), exhaustion (2.29/5, 0.78), and stress (1.95/5, 0.97)) showed no significant differences from the ratings before. This, combined with the rated demand of the experiment (3.57/5, 0.81), shows that the data should not be substantially distorted by tiredness effects. The post-questionnaire further showed that participants had a high motivation to follow the videos (mean=3.81/5, SD=1.08) and a very high motivation to perform well on the quizzes (mean=4.48/5, SD=0.75).

4.2 Content-Wise Ratings and Quiz Results

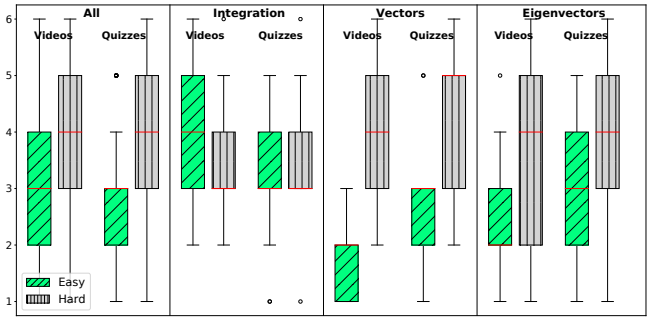
4.2.1 Content-wise Subjective Ratings. Figure 1 shows the CL and difficulty ratings for the quizzes and videos of each content item. While the differences across all topics, as well as within the topics vectors and eigenvectors, are clearly visible and significant (all $p < .01$), the integration content did not impose any statistically significant difference in perceived CL or difficulty.

4.2.2 Correlations of CL & Difficulty. CL and difficulty ratings correlate significantly (all $p < .01$) and strongly for all content items, with Pearson correlation coefficients between .58 (for the easy integration videos) and .89 (for the easy integration quiz). Thus, participants considered the two constructs as highly similar.

4.2.3 Quiz Time & Performance. Table 1 shows that strong differences in quiz times exist between the content items and that for all three topics, the quiz time was higher for the harder content. The average percentage of correct answers to the quizzes was lowest for the eigenvector quizzes, while the other two topics were comparable. On all three topics, students indeed performed better on quizzes



(a) Easy vs. hard CL rating per content.



(b) Easy vs. hard difficulty rating per content.

Figure 1: Subjective CL and difficulty ratings.

corresponding to simpler videos; however, the differences are very small, showing that such performance measures themselves do not always work as expected.

Table 1: Quiz time (in seconds) and performance (% of correct answers) mean and σ (in brackets) for the content items.

	Integration		Vector		Eigenvector	
	Easy	Hard	Easy	Hard	Easy	Hard
Time	88 (31)	129 (49)	119 (53)	199 (81)	93 (35)	174 (103)
Perf.	93 (18)	91 (15)	94 (16)	87 (20)	85 (18)	81 (24)

4.2.4 Discussion. Overall our participants had comparable backgrounds in the explored topics, and the data should not be substantially distorted by tiredness effects. As anticipated, there is a significant difference between easy and hard content for both videos and quizzes as well as CL and difficulty ratings except when considering the integration topic individually. Thus, we should also investigate each topic on its own. Furthermore, we see a strong correlation between CL and difficulty, indicating that participants perceive the two constructs as very related. While participants indeed required more time for the quizzes on hard content, the percentage of correct answers was rather comparable.

4.3 Predictive Models

We now aim to use the various captured measures to predict the demand imposed by the content as defined by our three measures

IntrinDiff, SubjDiff, and SubjCL. After looking at correlations between features and these measures, we investigate how to best select a subset of the implemented features and which models are most suitable for these classification and regression tasks. Then we train the actual models and discuss their respective results.

4.3.1 Correlation to Target Variables. First we analyze how strongly the individual features correlate to our target variables. Since we have many features and 3 target variables, we do not want to look at each individual correlation. Instead, we look at the highest correlations for both videos and quizzes, per topic and across topics.

Across videos, correlations are rather weak, where the maximum correlation coefficient of 0.2 was achieved for SubjCL (0.18 for SubjDiff and 0.14 for IntrinDiff). However, for the individual topics, we get much better results: for vectors, the best correlations are within 0.38 and 0.44 for the 3 target measures, for integration within 0.39 and 0.43, and for eigenvectors between 0.30 and 0.36. For the quizzes, correlations are much higher, both across all topics (between 0.38 and 0.42) and within topics, with the highest correlation coefficients between 0.48 and 0.52 for the vectors, 0.39 to 0.52 for integration, and 0.42 to 0.48 for eigenvectors.

4.3.2 Feature Amount & Model Selection. This section describes the experiments conducted to determine an appropriate model as well as an ideal number of features to use for training predictive models on our data. As a feature selection approach, we use recursive feature elimination with cross-validation (`RFECV` in `scikit-learn`) as it turned out to give better results than other feature selection approaches that we explored. As possible numbers of features, we test values ranging from 5 to 100 with an increment of 5. As for machine learning models, which also influence the number of features to select, we test the following models: linear models with different regularizers, namely a Logistic Regression, a Stochastic Gradient Descent regressor, a Lasso model, an Elastic Net, and a Ridge regressor, as well as a non-linear Random Forest regressor, all provided in the `scikit-learn` library. We further integrate linear mixed-effect models (LMEMs) using R (version 3.6.0, lme4 package version 1.1-21), as these can effectively capture inter-participant differences by adding a random effect for subject and/or content[3].

For each model and feature amount combination, we test different hyper-parameter settings of the model to get its best performance on that number of features[4]. Finally, we plot all models' performances for CL rating, difficulty rating, and intrinsic difficulty, once for videos and once for quizzes.

Across all 6 cases, we get comparable results: For the classification case (IntrinDiff), the best results were achieved using Logistic Regression or LMEM models, especially for a small number of features. For regression (SubjDiff, SubjCL), LMEM and Ridge performed best, showing that linear models seem to perform well on our data. Regarding feature amount, the range of 30 to 35 features gave good results across all 6 analyses. Since fewer features help interpret the results, we decided to use RFECV with 30 features in the following. While both linear models and LMEMs perform equally well in this preliminary analysis, Ridge and Logistic Regression are even simpler than LMEM, giving better generalization due to less participant dependence, which is why we use them in the remaining analyses.

To avoid overfitting, 10-fold cross-validation was used, and the best hyper-parameters determined by grid search, namely Ridge regression with $\alpha = 2$ for regression, and Logistic Regression using L2 normalization and $C = 1$, were chosen for the remaining analyses. As features, we use all features presented in Section 3.2, with a few exceptions: as the subjective measures are our target variables, we do not use them as predictors. Furthermore, we exclude the performance measures, as these exist only for the quizzes and not for the videos, resulting in a total of 202 features. Furthermore, if every entry for a whole feature contains the same value, we drop it (which happened for 3 "minimum" features). If due to a sensor failure some data values of a feature are missing, we replace them by the participant's mean value for that feature (if available), or by the global mean (if no data exists for a particular feature for that participant), which happened 5 times. Furthermore, we apply a z-transformation to achieve 0 mean and unit variance. For combining individual features within a modality or across modalities, we then use simple vector concatenation.

4.3.3 Classifying intrinsic difficulty. Using the settings described above, we train models classifying IntrinDiff, i.e., binary classification. Figure 2 depicts the accuracy achieved by the Logistic Regression models in comparison to a simple baseline always predicting 'easy' (achieving 50% accuracy). As can be seen, distinguishing easy from hard quizzes based on the sensor data works very well for the quizzes (80-90% accuracy), both across topics and within topics. For the videos, however, only around 70-75% accuracy is achieved for the vector and eigenvector topics as well as across topics. A reason could be that the videos are consumed only passively, where sensor data might be less reliable. This difference is also visible in the correlations above (Section 4.3.1), where higher coefficients were found for quizzes than for videos. For the integration videos, very high classification results were achieved, probably due to some artifact in the data, for which we currently do not have a concrete explanation. One should note here that these results were achieved using feature selection on all available features; therefore, Section 4.4 explores how results change when only single modalities are used.

4.3.4 Regression for predicting SubjDiff *and* SubjCL. Next, we check the MSE for the regression models predicting SubjDiff and SubjCL for both quizzes and videos in comparison to simple baselines always predicting the mean value of the corresponding rating, as depicted in Figure 3. For each topic individually, substantial percentage gains are achieved over the baseline; across quizzes, the prediction also yields good results, while across videos, only marginal gains were achieved. A potential explanation might be that the differences when passively watching videos are less well represented in the physiological data than those that appear when actively working on the quizzes, especially when adding the variability of the different topics instead of comparing content within

[3]Since the R package used for LMEMs does not support our feature selection approach, we instead perform feature selection with a Ridge model for regression and Logistic Regression for classification. For classification, we did not add a random effect for item (in our case, the video/quiz), as this would have trivially resulted in 100% accuracy.

[4]For SGD regressors, we explore L1 ratios of 0.15 and 0.5; for Lasso models alpha values 1, 2, and 10; for ElasticNet alpha values of 0.5 and 1 in combination with L1 ratios of 0.25, 0.5, 0.75; for Random Forests (both for regression and classification) we explore 10, 20, 30, and 50 for numbers of estimators and a maximum depth of None, 4, 8, and 12. For Ridge models, we explored alpha values of 0.5, 1, 2, and 10; for SGD classifiers, alpha values 0.0001, and 0.01, with L1 and L2 regularization; for Logistic Regression C values of 0.5, 1, and 2, both with L1 and L2 regularization.

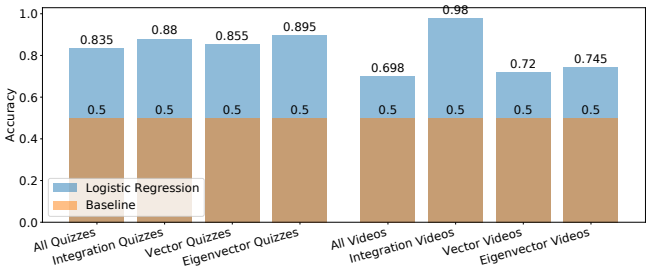
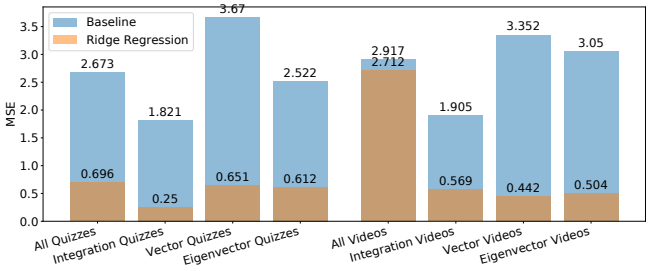
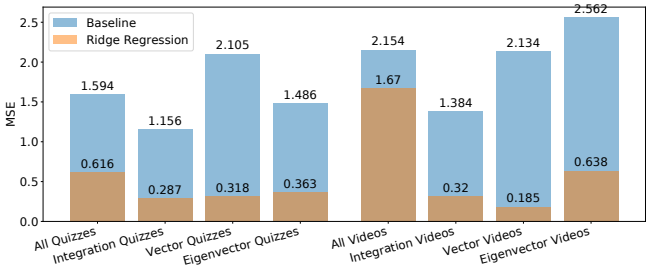


Figure 2: Classification of intrinsic difficulty.

a topic. Section 4.3.1 also shows that the correlations across videos are much lower than within topics or across quizzes, explaining the bad results in this particular case. It is also interesting that the prediction of SubjCL and SubjDiff also works for the integration topic, where the subjective differences were not large, which again can be explained by the existing correlations presented in Section 4.3.1. Overall, the final MSEs found are very low (except for the across videos-case), indicating that within content topics, one can estimate the imposed demand very well.



(a) CL



(b) Difficulty

Figure 3: Regression performance for SubjCL **and** SubjDiff**.**

4.3.5 Discussion. Overall, this initial analysis shows that using roughly 30 features together with Ridge/Logistic Regression is a reasonable choice for our dataset. Furthermore, for both the classification and regression case, results for the quizzes are better than for the videos, which we believe comes from the higher degree of activity while solving quizzes than while watching videos, which in turn better differentiates the physiological data. Therefore, using feature selection on all modalities simultaneously proves to work very well for the quizzes for binary classification of IntrinDiff and for all regression cases except for predicting across videos. However, note that the limited amount of data might have introduced some bias, even though the results look consistent.

4.4 Modality Analysis

In this section, we aim to understand which feature modalities contribute how strongly to the models.

4.4.1 Modality Correlations to Target Variables. To estimate the direct link between the modalities and the 3 target variables, we analyze the 10 highest correlating features within the 8 cases (across quizzes/videos, within each quiz/video). Of these in total 240 (3*10*8) features, 113 are eye features, 68 heart features, 53 skin features, and 6 body posture features, suggesting that eye features perform best, followed by heart and skin measures.

Naturally, this approach has some limitations: the same feature could count up to 24 times (to all target variables of all 8 cases), and there is a different number of features per modality. However, this initial analysis captures the direct link to the target measures independent of there being even more irrelevant features in the modality, and independent of linear dependency and thereby redundancy of multiple features. Nevertheless, to also investigate the predictive power, further analyses are presented in the following.

4.4.2 Modalities Selected through Feature Selection. We analyze the features selected among all possible features to see if measures from some modality tend to be selected more or less often by our feature selection approach. We compare this among the total of 24 tasks (3 target measures times the 8 cases: across videos/quizzes and within the 3 topics each with video/quiz) for which we train our models. To better analyze the selected feature set, we count which individual features are selected most often. Then we check which modalities these highly selected features belong to. This shows that eye and heart features are selected most often (up to 17 out of 24 times), showing the importance of these features. In contrast, skin and body posture features are maximally selected 7 times, and can therefore be considered less important.

4.4.3 Intraclass Correlation Coefficient. It is also interesting to analyze the degree to which participants resemble each other regarding a given feature. For this, we use the Intraclass Correlation Coefficient (ICC). It ranges between 0 (chance agreement) and 1 (perfect agreement) and gives "an indication of the extent to which different [participants] produce the same values of a given measure when exposed to the same [conditions]" [60]. It therefore indicates features that already generalize well from a small number of participants. According to Koo et al. [29], values below 0.5 can be considered as 'poor', values between 0.5 and 0.75 as 'moderate', values between 0.75 and 0.9 as 'good', and everything above 0.9 as 'excellent'.

Overall, only 2.5% of the features can be considered 'excellent', 3.5% as 'good', 27% as 'moderate', and the majority of 67% as 'poor'. This is particularly interesting because the investigated measures were all proposed in the literature and used in CL studies, which usually do not report the ICC and mostly have a similar or even smaller amount of participants. Comparing the values to one of the few papers that also reported ICC values on multiple modalities [60], we find that their 7 explored features were all in the range 0.25 to 0.6, and thus also 'poor' to 'moderate' according to Koo et al. [29].

The high amount of features in the 'poor' category can also be explained by the fact that not all of the 5 features calculated on top of continuous signals (min, max, avg, sd, range) consistently yield good ICC values. Furthermore, it can be seen as an explanation for why selecting 30 features already gives good results.

In terms of modalities, again, eye features perform best, comprising 100% of the 'excellent', 57% of the 'good', and 11% of the 'moderate' features. Between heart and skin, there does not seem to be a clear winner: heart features make up only 14% of the 'good' but 48% of the 'moderate' features, whereas skin yields 29% of the 'good' and only 11% of the 'moderate' features. All body posture features are within the 'poor' category.

4.4.4 Prediction Performance of Different Modalities. To get more insight into the classification/regression performance that can be achieved through the different modalities (heart, skin, eyes) and combinations thereof, we train models on the combined modalities and subsets of the modalities. We ignore body posture features here, as the performance was very poor.

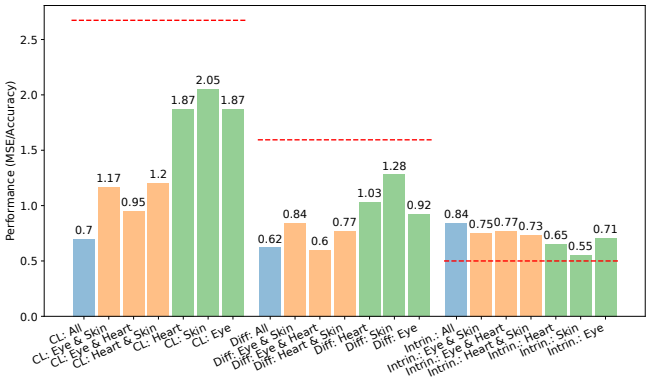


Figure 4: Regression (SubjDiff **and** SubjCL**) and classification performance (**IntrinDiff**) for prediction across all quiz contents when using only features from certain modalities.**

Figure 4 shows the results achieved compared to the baseline reported above, each plot containing a group for SubjCL, for SubjDiff (both MSE), and for IntrinDiff (as accuracy). Note that this analysis was done across all quiz contents; the corresponding analysis across all video contents is omitted, because using all modalities across videos already resulted in only marginal gains compared to the baseline (see Figure 3). The resulting plot therefore did not provide further insights and is omitted for space reasons.

As can be seen in the figure, the multi-modal approach is consistently better than single modalities; however, the combination of eye and heart features is also comparably good. Furthermore, we note that there is a trend that heart and eye features perform better than skin features, which can also explain why the combination of the two outperforms the modality pairs containing skin.

4.4.5 Discussion. The various sub-analyses conducted to analyze which modalities perform better and worse, some focusing on correlations while others focused on predictive power, consistently show that eye features perform best, followed by heart, then skin, and last body posture. Combining two modalities improves results compared to single modalities, where eye and heart features combined performed best. We thereby extend the findings by Naismith and Cavalcanti [40], who showed that eye features are more reliable than cardiovascular features in medical training, by additionally considering skin features and combinations of modalities.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present an approach based on a wide range of physiological, behavioral, performance, and subjective measures, yielding the so far most diverse set of features from a variety of modalities that has been investigated for estimating the demand imposed by e-learning content. Our study with 21 participants used a rather realistic e-learning setting, where participants learn through videos and quizzes. With the data captured in this setting, we show that classifying intrinsic content difficulty works better for quizzes, where participants actively solve problems, than for videos, which they passively consume. Our classification results are roughly comparable to [6], which achieved up to 73% in a trinary classification task (low CL, high CL, without task), but used a much less realistic setting where participants had to perform mental calculations instead of learning through videos and quizzes. It is also interesting that even though we did not use EEG measures, the combined power of multiple modalities gives comparable results. Regression analysis for predicting the subjectively reported level of CL and difficulty also works with very low error within content topics. Among the explored feature modalities, eye-based features yield the best results, followed by heart-based and then skin-based measures. Furthermore, combining multiple modalities results in better performance compared to using a single modality. The presented results can guide researchers and developers of cognition-aware e-learning environments by suggesting modalities and features that work particularly well for estimating difficulty and CL. Furthermore, the results suggest that adaptations like content recommendations, break proposals, or speed adaptations would be feasible using a multi-modal approach. One should however note that the data was captured from only 21 participants learning 6 mathematical contents, so further studies should be conducted in different domains with more participants.

Currently, all our features are calculated on all data available per content item, which is sufficient to predict perceived CL after finishing that content, i.e., the average load [65]. However, as discussed in Section 3.2.5, all of our continuous features could also be calculated on smaller time intervals, which would allow the system to quickly adapt to changes in the user's states as proposed in [47], and would provide further insights into which modalities work well for this real-time setting. Apart from this analysis, we plan to gain further insights into the reliability of individual features as opposed to the modality level analysis that we presented here for space reasons. Last, we aim to develop similar cognition-aware systems in other demanding domains such as translation [18].

ACKNOWLEDGMENTS

This research was funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS17043 (project KOALA). Many thanks to Carsten Müssig and Erik Schulze for their valuable feedback and suggestions.

REFERENCES

[1] Syed Arshad, Yang Wang, and Fang Chen. 2013. Analysing mouse activity for cognitive load detection. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*. ACM, 115–118.

[2] Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. 2009. Estimation of behavioral user state based on eye gaze and head pose – application in an e-learning environment. *Multimedia Tools and Applications* 41, 3 (2009), 469–493.

[3] Alan D Baddeley and Robert H Logie. 1999. Working memory: The multiple-component model. (1999).

[4] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments* 24, 3 (2016), 590–605.

[5] Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods* 190, 1 (2010), 80–91.

[6] Magdalena Borys, Małgorzata Plechawska-Wójcik, Martyna Wawrzyk, and Kinga Wesołowska. 2017. Classifying cognitive workload using eye activity and EEG features in arithmetic tasks. In *International Conference on Information and Software Technologies*. Springer, 90–105.

[7] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2008. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*. ACM, 13–17.

[8] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust Multimodal Cognitive Load Measurement*. Springer.

[9] Jingjing Chen. 2016. *Enhancing Student Engagement and Interaction in E-Learning Environments through Learning Analytics and Wearable Sensing*. Ph.D. Dissertation.

[10] Siyuan Chen and Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine* 110, 2 (2013), 111–124.

[11] Vera Demberg and Asad Sayeed. 2016. The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PloS One* 11, 1 (2016), 1–29.

[12] Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 24, 1 (2010), 1–13.

[13] Federico Cirett Galán and Carole R Beal. 2012. EEG estimates of engagement and cognitive workload predict math problem solving outcomes. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 51–62.

[14] Giuseppe Ghiani, Marco Manca, and Fabio Paternò. 2015. Dynamic user interface adaptation driven by physiological parameters to support learning. In *Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, 158–163.

[15] Joseph H Goldberg and Xerxes P Kotval. 1999. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics* 24, 6 (1999), 631–645.

[16] Markus Guhe, Wayne D Gray, Michael J Schoelles, Wenhui Liao, Zhiwei Zhu, and Qiang Ji. 2005. Non-intrusive measurement of workload in real-time. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications Sage CA: Los Angeles, CA, 1157–1161.

[17] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 301–310.

[18] Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Multi-Modal Approaches for Post-Editing Machine Translation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 231.

[19] Nico Herbig, Santanu Pal, Mihaela Vela, Antonio Krüger, and Josef van Genabith. 2019. Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation* (2019), 1–25.

[20] Nico Herbig, Patrick Schuck, and Antonio Krüger. 2019. User acceptance of cognition-aware e-learning: An online survey. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 17.

[21] G Robert J Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology* 45, 1 (1997), 73–93.

[22] Gahangir Hossain and Mohammed Yeasin. 2014. Understanding effects of cognitive load from pupillary responses using Hilbert analytic phase. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 375–380.

[23] Seyyed Abed Hosseini and Mohammad Ali Khalilzadeh. 2010. Emotional stress recognition system using EEG and psychophysiological signals: Using new labelling process of EEG signals in emotional stress state. In *International Conference on Biomedical Engineering and Computer Science*. IEEE, 1–6.

[24] Cristina Iani, Daniel Gopher, and Peretz Lavie. 2004. Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology* 41, 5 (2004), 789–798.

[25] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *Extended Abstracts on Human Factors in Computing Systems*. ACM, 1477–1480.

[26] Shoya Ishimaru, Soumy Jacob, Apurba Roy, Syed Saqib Bukhari, Carina Heisel, Nicolas Großmann, Michael Thees, Jochen Kuhn, and Andreas Dengel. 2017. Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 8. IEEE, 32–36.

[27] Slava Kalyuga, Paul Chandler, and John Sweller. 1998. Levels of expertise and instructional design. *Human Factors* 40, 1 (1998), 1–17.

[28] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. 2012. Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *Journal of Physical Therapy Science* 24, 12 (2012), 1341–1344.

[29] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163.

[30] Arthur F Kramer. 1991. Physiological metrics of mental workload: A review of recent progress. *Multiple-Task Performance* (1991), 279–328.

[31] Fan-Ray Kuo, Gwo-Jen Hwang, Yen-Jung Chen, and Shu-Ling Wang. 2007. Standards and tools for context-aware ubiquitous learning. In *Proceedings of the Conference on Advanced Learning Technologies*. 704–705.

[32] Derick Leony, Abelardo Pardo Sánchez, Hugo A Parada Gélvez, and Carlos Delgado Kloos. 2012. A widget to recommend learning resources based on the learner affective state. In *CEUR-Workshop Proceedings*.

[33] Jimmie Leppink. 2017. Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences* 12, 5 (2017), 385–391.

[34] Accuray Research LLP. 2017. *Global E-Learning Market Analysis & Trends – Industry Forecast to 2025*. https://www.researchandmarkets.com/reports/4039818/global-e-learning-market-analysis-and-trends

[35] Yu Lu, Sen Zhang, Zhiqiang Zhang, Wendong Xiao, and Shengquan Yu. 2017. A framework for learning analytics using commodity wearable devices. *Sensors* 17, 6 (2017), 1382.

[36] Philipp Mock, Peter Gerjets, Maike Tibus, Ulrich Trautwein, Korbinian Möller, and Wolfgang Rosenstiel. 2016. Using touchscreen interaction data to predict cognitive workload. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 349–356.

[37] Barbara Moissa, Geoffray Bonnin, and Anne Boyer. 2019. Exploiting wearable technologies to measure and predict students' effort. In *Perspectives on Wearable Enhanced Learning (WELL)*. Springer, 411–431.

[38] Joss Moorkens, Sharon O'Brien, Igor AL da Silva, Norma B de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29, 3-4 (2015), 267–284.

[39] LJM Mulder. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology* 34, 2 (1992), 205–236.

[40] Laura M Naismith and Rodrigo B Cavalcanti. 2015. Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine* 90, 11 (2015), S24–S35.

[41] Sharon O'Brien. 2006. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology* 14, 3 (2006), 185–205.

[42] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* 38, 1 (2003), 63–71.

[43] Fred GWC Paas and Jeroen JG Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review* 6, 4 (1994), 351–371.

[44] Ana M Pernas, Adenauer C Yamin, João LB Lopes, and Jose P M de Oliveira. 2014. A semantic approach for learning situation detection. In *Proceedings of the Conference on Advanced Information Networking and Applications*. 1119–1126.

[45] Manuel Rodrigues, Sérgio Gonçalves, Davide Carneiro, Paulo Novais, and Florentino Fdez-Riverola. 2013. Keystrokes and clicks: Measuring stress on e-learning students. In *Management Intelligent Systems*. Springer, 119–126.

[46] Dennis W Rowe, John Sibert, and Don Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the Conference on Human Factors in Computing Systems*. 480–487.

[47] Holger Schultheis and Anthony Jameson. 2004. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, 225–234.

[48] Fred Shaffer and JP Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in Public Health* 5 (2017), 258.

[49] Liping Shen, Victor Callaghan, and Ruimin Shen. 2008. Affective e-learning in residential and pervasive computing environments. *Information Systems Frontiers* 10, 4 (2008), 461–472.

[50] Liping Shen, Minjuan Wang, and Ruimin Shen. 2009. Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal*

of Educational Technology & Society 12, 2 (2009).

[51] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Extended Abstracts on Human Factors in Computing Systems*. 2651–2656.

[52] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. 2012. Brainput: Enhancing interactive systems with streaming fNIRS brain input. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2193–2202.

[53] T Soukupova and Jan Cech. 2016. Real-time eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop*. 1–8.

[54] Martin Spüler, Carina Walter, Wolfgang Rosenstiel, Peter Gerjets, Korbinian Moeller, and Elise Klein. 2016. EEG-based prediction of cognitive workload induced by arithmetic: A step towards online adaptation in numerical learning. *ZDM* 48, 3 (2016), 267–278.

[55] Els Stuyven, Koen Van der Goten, André Vandierendonck, Kristl Claeys, and Luc Crevits. 2000. The effect of cognitive load on saccadic eye movements. *Acta Psychologica* 104, 1 (2000), 69–85.

[56] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10, 3 (1998), 251–296.

[57] Marten van den Berg, Peter Rijnbeek, Maartje Niemeijer, Albert Hofman, Gerard van Herpen, Michiel Bots, Hans Hillege, Kees Swenne, Mark Eijgelsheim, Bruno Stricker, et al. 2018. Normal values of corrected heart-rate variability in 10-second electrocardiograms for all ages. *Frontiers in physiology* 9 (2018), 424.

[58] Karl F Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye activity correlates of workload during a visuospatial memory task. *Human Factors* 43, 1 (2001), 111–121.

[59] Philip A Vernon. 1993. Der Zahlen-Verbindungs-Test and other trail-making correlates of general intelligence. *Personality and Individual Differences* 14, 1 (1993), 35–40.

[60] Lucas Nunes Vieira. 2016. How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. *Machine Translation* 30, 1-2 (2016), 41–62.

[61] María Viqueira Villarejo, Begoña García Zapirain, and Amaia Méndez Zorrilla. 2012. A stress sensor based on Galvanic Skin Response (GSR) controlled by ZigBee. *Sensors* 12, 5 (2012), 6075–6101.

[62] Susanne Vogel and Lars Schwabe. 2016. Learning and memory under stress: Implications for the classroom. *npj Science of Learning* 1 (2016), 16011.

[63] Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, Peter Gerjets, and Martin Spüler. 2017. Online EEG-based workload adaptation of an arithmetic learning environment. *Frontiers in Human Neuroscience* 11 (2017), 286.

[64] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang. 2016. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology* 47, 6 (2016), 1304–1323.

[65] Bin Xie and Gavriel Salvendy. 2000. Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & Stress* 14, 1 (2000), 74–99.

[66] Takehiro Yamakoshi, K Yamakoshi, S Tanaka, M Nogawa, Sang-Bum Park, Mariko Shibata, Y Sawada, P Rolfe, and Yasuo Hirose. 2008. Feasibility study on driver's stress detection from differential skin temperature measurement. In *Engineering in Medicine and Biology Society*. IEEE, 1076–1079.

[67] Zhiwen Yu, Yuichi Nakamura, Seiie Jang, Shoji Kajita, and Kenji Mase. 2007. Ontology-based semantic recommendation for context-aware e-learning. In *International Conference on Ubiquitous Intelligence and Computing*. Springer, 898–907.