Multimodal data as a means to understand the learning experience

Michail N. Giannakos[a,*], Kshitij Sharma[a], Ilias O. Pappas[a,c], Vassilis Kostakos[b], Eduardo Velloso[b]

[a] Department of Computer and Information Science, Norwegian University of Science and, Technology (NTNU), Sem Sælands vei 7-9, 7491, Trondheim, Norway
[b] School of Computing & Information Systems, University of Melbourne, Room 9.03, Doug McDonell Building 168, VIC-3010 Parkville, Australia
[c] University of Agder (UiA), Kristiansand, Norway

ARTICLE INFO

ABSTRACT

Most work in the design of learning technology uses click-streams as their primary data source for modelling & predicting learning behaviour. In this paper we set out to quantify what, if any, advantages do physiological sensing techniques provide for the design of learning technologies. We conducted a lab study with 251 game sessions and 17 users focusing on skill development (i.e., user's ability to master complex tasks). We collected click-stream data, as well as eye-tracking, electroencephalography (EEG), video, and wristband data during the experiment. Our analysis shows that traditional click-stream models achieve 39% error rate in predicting learning performance (and 18% when we perform feature selection), while for fused multimodal the error drops up to 6%. Our work highlights the limitations of standalone click-stream models, and quantifies the expected benefits of using a variety of multimodal data coming from physiological sensing. Our findings help shape the future of learning technology research by pointing out the substantial benefits of physiological sensing.

1. Introduction

We present a study that measures how well traditional click-stream models are associated with human learning, and contrast those against multimodal data-stream models based on from physiological responses. We consider physiological data captured during user interaction with technologies that support learning (i.e., learner–computer interaction – LCI), and we focus on skill acquisition. Our work seeks to provide insights on how multimodal data captured via physiological sensing can help us improve user interfaces for learning, and is a first step towards critically designing to amplify human learning.

The evidence for understanding and supporting users' learning experience is still very limited, considering the wide range of multimodal data produced when the learner interacts with a system (Giannakos et al., 2018). Cameras, wearable sensors, biosensors, infrared imaging, eye-tracking and more, offer the opportunity to enhance the way we collect and analyze user-data to achieve a deeper understanding of the interaction between humans and learning technologies (Ochoa and Worsley, 2016; Blikstein and Worsley, 2016). Despite the great potential of multimodal data – and in particular physiological data – for understanding users' cognition, emotion, attention, information acquisition and more (Zheng et al., 2018; Kanjo et al., 2015; Hassib et al., 2016; Scheidt and Chung, 2018; Abdelrahman et al, 2017), research in

this direction remains scarce. Today, we do not make full use of potential analytics, since we only utilize part of the available user-data (e.g., click-streams, surveys, preferences) in learning technology (based on a recent literature review in learning analytics for learning design (Mangaroska and Giannakos, 2018)). At the same time, the advances in learning technologies (Chang, 2016; Yousafzai et al., 2016; Rana et al., 2016; Rana and Dwivedi, 2018; Panigrahi et al., 2018) create an increasingly more challenging context for learning, by consuming significant cognitive resources (e.g., attention), by frequent task changes (i.e., lack of focus), by an ever-increasing information overload, by generating unwanted distractions (i.e., various notifications), and often including subpar user interfaces, dashboards and communication modalities (Niforatos et al., 2017).

Learning with technology is a complex process that is associated with many aspects of interaction (Di Mitri et al., 2017) (e.g., hard mental operations, cognitive friction, etc.). The complexity of this process means that it is likely that no single data modality can paint a complete picture of the learning experience, requiring multiple data streams to complement each other (Zheng et al., 2018). Integrating this information with fusion technologies is attractive for building robust user and learner models (Di Mitri et al., 2017) as well as for developing a holistic understanding of learner-computer interaction (D'mello and Kory, 2015). However, most studies have focused on using one data-

stream (mainly user actions/clicks) or combining limited data-streams (e.g., user actions/clicks with self-reported data) for measuring learning experience. In contrast, the combination of physiological data coming from the central nervous system (i.e. electroencephalography – EEG) and external behaviors (i.e. eye-movements) has been reported as a promising approach (Zheng et al., 2018). Recent studies have shown that combining multimodal data-streams significantly increases the accuracy and generates more insights about the user experience (Zheng et al., 2018). For example, brain and eye movement signals convey important information about users' social and emotional information for context-aware environments (Lee and Anderson, 2017). Thus, a research challenge is to harness multimodal user-data and utilize data-driven insights to design technologies that support human learning.

In this paper, we present a lab study in which we capture data produced during interaction with a game that focuses on simple skill development (i.e., intuitive learning/mastery through play (Lai et al., 2018)). We collected data coming from multiple sources (i.e. click-stream, eye-tracking, EEG, video, and wristband). We applied machine learning techniques to extract and select the features associated with skill acquisition. By identifying those features, we provide a path towards the design of systems that better support learning.

The paper provides new insights on the role of multimodal data in technology enhanced learning. In particular, we make the following contributions:

- We present insights from a controlled experiment that collected five different data-streams during a basic user/learner–computer interaction task.
- We show that multimodal data-streams have the capacity to give more accurate prediction of users' skills acquisition compared to traditional click-stream models.
- We identify the physiological features that best predict skill development.
- We discuss how our findings can be used to design future learning technologies, as well as to advance research in the area of learning technologies.

The paper is structured as follows. The next section outlines the background and related work for this study. The third section outlines the employed methodology. The forth section presents the results of the study. The fifth section of the paper, discuss the results, the limitations, and the implications of the study, and the last section provides the conclusions and the future work.

2. Background and related work

2.1. Click-streams capacity to predict learning performance

Earlier studies have utilized various performance-prediction techniques in different types of click-stream and keystroke data (Ashenafi et al., 2016; Watson et al., 2013). Early works focused on predicting the overall completion and outcome such as dropping out or completing, or passing or failing a course (Manh aes et al., 2014), while more recent fine-grained and sophisticated approaches involved predicting actual scores for different tasks like tests and assignments (Elbadrawy et al., 2016; Moreno-Marcos et al., 2018).

Due to the diversity of the tasks in which such experiments have been conducted, the different interactions and global data collected, as well as the various advanced machine learning techniques that were applied (Ashenafi et al., 2016), it is difficult to compare results across these studies. Previous works have collected large amounts of global data about learners' activity with the learning system, demographics, self-reports and previous performance of the learner (Moreno-Marcos et al., 2018). Examples of studies that utilize information from students' activities in various learning technologies (e.g., LMSs, MOOCs, assessment platforms) in predicting performance have demonstrated the predictive power of these data-streams (Maldonado-Mahauad et al., 2018), as well as the importance of fusing those data-streams to increase their predictive power (Katerina Mangaroska and Giannakos, 2019).

So far we have seen several studies utilizing click-stream and keystroke data to predict learners' success (Moreno-Marcos et al., 2019). Ashenafi et al. (2016) used interaction data from a peer-assessment environment and combined them with the difficulty, interestingness and relevance of the tasks as indicated from the students, to build learner progress prediction models with an error rate of 12%. In the context of an online course about Java programming, Moreno-Marcos et al., (Moreno-Marcos et al., 2018) utilized the interaction data collected by the EdX platform (EdX, 2018) and found that forum-related variables do not add power to predict grades, but information about the previous performance does. Furthermore, the type of task can vary the results. In their study, with combined EdX interaction data (EdX, 2018) and information about the previous performance their prediction models indicated an error rate of 14% (Moreno-Marcos et al., 2018). Maldonado-Mahauad et al. (2018) utilized interaction data with the learning materials and the digital assessment as well as self-reported data (i.e., learning strategies), to develop prediction models of learners' grade. Their results indicate an average of 15% error rate and the best cases with 11.3% error rate. In another study (Elbadrawy et al., 2016), the authors used again activity data, course information and LMS data-features and obtained predictions with 13.5% error rate. In the same vein, Ren et al. (2016) developed a model to predict the grade for a student, prior to attempting the digital assessment. They developed a real-time model that was tracking the participation of a student within an online course (via click-stream server logs) and predicting his/her performance on the next assessment (within the course offering). The optimal performance of their model had an error rate of 11%. Overall we see that there is a wide variety in the choice of prediction features, but clickstream data use stands out (Moreno-Marcos et al., 2019). Clickstream and keystroke data demonstrate an error rate of around 20% (Moreno-Marcos et al., 2018), and the majority of the studies utilize other available data (e.g., self-reported, previous performance, demographics etc.) leading to better performance (error rate around 15%, with exceptional cases having an error rate of 11–12%).

Despite the demonstrated predictive power of the models based on the aforementioned data, they have some limitations. For example, frequency counts of events from clickstream data and other clickstream traces that are obtained directly from low-level data are limited for detecting learners more complex behaviour (Maldonado-Mahauad et al., 2018). Moreover, as previous studies already demonstrated, clickstreams in isolation do not necessarily build very accurate predictive models (Zhao et al., 2016). Therefore, predictive models could be improved by adding variables that encapsulate complimentary information or are build on longer activity sequences resulting from learners' interaction with the technology (Maldonado-Mahauad et al., 2018). Thus, this paper builds on prior work on MultiModal Learning Analytics (MMLA) and attempts to shed light on the capacity of learner-generated data, beyond clickstreams, to explain learning.

2.2. Multimodal data and learning experience

Interaction between learners and technologies is complex and offers an opportunity for collecting rich and multimodal data (Giannakos et al., 2018). Overcoming the difficulties in gathering and making sense of such data holds the potential to offer novel principles to support learning experiences (Ochoa and Worsley, 2016). Typical examples of such multimodal data (or MultiModal Learning Analytics – MMLA, as literature refers to them) include audio, video, electrodermal activity data, eye-tracking, user logs and click-stream data to name a few (Blikstein and Worsley, 2016; Prieto et al., 2018).

Several approaches exist to collect and analyze multimodal data during LCI, but they are not yet widespread and are largely limited to

mainstream data (e.g., keystrokes, click-streams). In the context of university lectures, researchers have used computer vision (Raca et al., 2015) to measure student attention based on their body language. Ochoa et al. (2013) gathered user data from video, audio and pen strokes to extract simple features that can help identify students with certain expertise in the area of study. More recently, a low cost system called RAP was proposed for collecting data from students' posture, gaze, volume of their voice and the content used, in order to generate a feedback report with multimodal recordings of students' performance (Ochoa et al., 2018). Another recent study (Prieto et al., 2018) collected eye-tracking, audiovisual and accelerometer data of the teacher in order to extract orchestration graphs (i.e., teaching activities and their social plane over time). We have also seen examples of multimodal data in the area of professional development, as Martinez et al. (Martinez-Maldonado et al., 2017; Echeverria et al., 2018) used sensor data (coming from patient manikins used for healthcare training) to capture students' interaction traces and identify the key aspects (e.g., instructor–student dynamics and movements) of the learning process. Existing studies utilize various user-generated data, including click-streams (Giannakos et al., 2015; Wang et al., 2016), log data (Pardo et al., 2017) and sensor data (Prieto et al., 2018) to inform visual representations (e.g., dashboards) and provide explicit, easy to understand, and concise ways of presenting information for improved LCI and informed decision making. However, there are limited insights on how such data can portray learning experience and what are the features that are associated with learning.

Evidence for understanding how human learning occurs during the interaction with the learning technology is still very limited, considering the abundance of user-generated data. The seamless integration and combination of different apparatuses, as well as the harmonization and sense-making of multimodal data-streams to support the learning experience, is an extremely challenging process (Lahat et al., 2015). However, with the recent technological developments in high-frequency data collection there is an unparalleled opportunity to understand how humans learn with technology and to use these insights to design systems that amplify human learning (Blikstein and Worsley, 2016). Cameras, wearable sensors, biosensors, gesture sensing, eye-tracking, among others, can help us to enhance the way we collect and make-sense of user-data to provide a deeper understanding of the interaction between humans and technologies (Ochoa and Worsley, 2016).

2.3. Physiological data

In recent years, *action-based analytics* have been found to be very promising in portraying the user experience during learning (Mangaroska and Giannakos, 2018; Blikstein and Worsley, 2016; Pardo et al., 2017; Rana and Dwivedi, 2016). When a user performs an action, regardless of whether it is completed or not, they produce rich information that is often not considered in designing technologies to support learning (e.g., brain-activity, eye-activity, facial gesture). These physiological-analytics can provide meaningful insights (e.g., users attention and cognitive state) to advance learning experience research (Pantazos and Vatrapu, 2016). For example, the main EEG frequencies of humans are categorized into four wave patterns: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz) and Beta (13–30 Hz) (Teplan et al., 2002). Research has shown that these patterns are strongly correlated to human emotions and cognitive states (Szafir and Mutlu, 2013; Hassib et al., 2017a,b), and are widely used to accurately estimate task engagement and cognitive load based on the amplitudes of Alpha, Beta and Theta waves. Wristband data, like electrodermal activity, have also been proved very valuable in monitoring continuously biophysical functions that have the capacity to accurately infer arousal levels (Exler et al., 2016; Niforatos et al., 2017).

Recent technological developments and data science advancements (Gandomi and Haider, 2015; Tursunbayeva et al., 2018) have boosted

the growth of non-invasive high-frequency physiological-data collections (e.g., rapid development of wearable devices, dry electrode techniques etc.), acquired through devices such as cameras, wearable sensors, biosensors (measuring skin conductivity, heartbeat, and brain activity), infrared imaging, and eye-tracking. However, to fully utilize these multi-faceted user-generated data, research needs to rigorously collect and harmonize them. Despite the great potential of physiological analytics in understanding users' senses, emotions, attention, information acquisition and more (Blikstein and Worsley, 2016; Ochoa and Worsley, 2016; Maior et al., 2018) and early results to inform designs and intuitive experiences (Feit et al., 2017; Bosch et al., 2015), their potential remains underexplored and underutilized. Thus, we propose that Human–Computer interaction (HCI) research should pursue multi-pronged approaches and the collection / combination of complex and multimodal data, as well as the identification of the most important features of those data-streams and their respective design implications.

2.4. Multimodal data collection and analysis

Utilizing representative, objective, diverse and accurate data allows us to better understand users and design meaningful experiences for them (Giannakos et al., 2018). Gaze has long been studied as an approach for understanding users' behaviors and cognitive states in HCI (Duchowski, 2007). For instance, pupil response has been successfully used to infer cognitive and emotional processes (Bradley et al., 2008), and pupil size has been used to infer cognitive listening load, user information processing and understanding (Zekveld et al., 2014). Other gaze measurements, such as fixations, saccades, and blinks, provide important cues for context-aware environments which reveal how a user interacts with their surroundings and what attracts a user's attention (Bulling et al., 2011). Brain data using EEG signals have also been used in HCI to detect shifts in engagement and workload. Properties of EEG signals such as the different frequency bands provide cognitive information with a high temporal resolution that can be linked to real-world stimuli (Hassib et al., 2017a,b). Despite the relatively expensive devices and specialized data analysis techniques, HCI research recently showed the feasibility of using consumer EEG sensors for sensing users' engagement and providing useful design insights in several domains (Huang et al., 2014; Hassib et al., 2017a,b; Mills et al., 2017). Thus, regardless of the difficulty in collecting and analyzing multimodal data generated during users' interaction with a computer, it is arguable in HCI community that such data have the capacity to provide rich information to inform design (what we call in this paper, *design capacities* of multimodal data).

Despite the promising and successful applications of multimodal data collections and the existence of the required statistical techniques, the capacities of multimodal data as a means to understand and improve LCI remain largely unexplored, since to the best of our knowledge there are no initiatives combining high variety of different multimodal data collections (i.e., EEG, eye-movements, video, keystrokes and wristband data). To capitalise on the design capabilities of multimodal data, research needs to simultaneously study them, such that their respective usefulness can be reliably assessed. This is the objective of our study, which we describe next.

3. Methods

3.1. Context

To collect a fine-grained multimodal dataset during LCI, and to be able to extract the features associated with learning, we designed a controlled experiment focusing on skill acquisition. Skill acquisition (commonly termed also as movement-motor learning (Ericsson et al., 2018)) is a loosely defined term that encompasses motor adaptation and decision-making (Wolpert et al., 2011; Krakauer and Mazzoni, 2011). Based on the cognitive approach, developing a skill leads to the

Fig. 1. Screenshot of the home made Pac-Man game, basic concept for the game, minimalistic design to achieve non-distracting environment.

acquisition and memorization of an internal representation of a movement (conceptualized as a motor schema) (Thon, 2015). Furthermore, learning such a skill allows to develop a motor program while defining at the same time a sensory reference of the movement before it is performed. To increase the efficiency of the sensory–motor loops that take part in movement control, we can process sensory feedback and knowledge of results. Sensory feedback refers to movement-related information and knowledge of results to movement outcomes. Finally, when learners receive guidance verbally or they rehearse mentally the skill to be learned, thus requiring cognitive effort, higher cognitive processes occur during learning (Thon, 2015).

Because we wanted to maintain a simple learning curve for the LCI task, we developed a time-tested game that has been used to measure specific skills (motor skills) in the past (Nicolson and Fawcett, 2000). Thus, we developed a Pac-Man game following all the game play elements and giving 3 lives for each session (see Fig. 1). The game was controlled by the 4 arrow buttons of the keyboard and was developed to log every keystroke of the user. The difficulty of the game increased from one session to another.

3.2. Participants

We recruited a total of 17 healthy participants (7 females) aged between 17 and 49 years (mean = 32.05, SD = 8.84). Participants were recruited from the participant pool of a major European university. Participants were familiar with the game, but none of them had played the game in the previous 2 years. Prior to completing the tasks, participants were informed of the purpose and procedure of the experiment and of the harmlessness of the equipment. Participants were given a movie theater ticket upon completion of the study.

3.3. Procedure

Upon obtaining consent, the researcher escorted the participant to the user experience (UX) room, which contained a chair facing a large computer monitor (see Fig. 2). The participant wore the wristband and EEG cap, and connected and then calibrated all the data collection devices (i.e., eye-tracker, wristband, EEG, cameras). The eye-tracking,
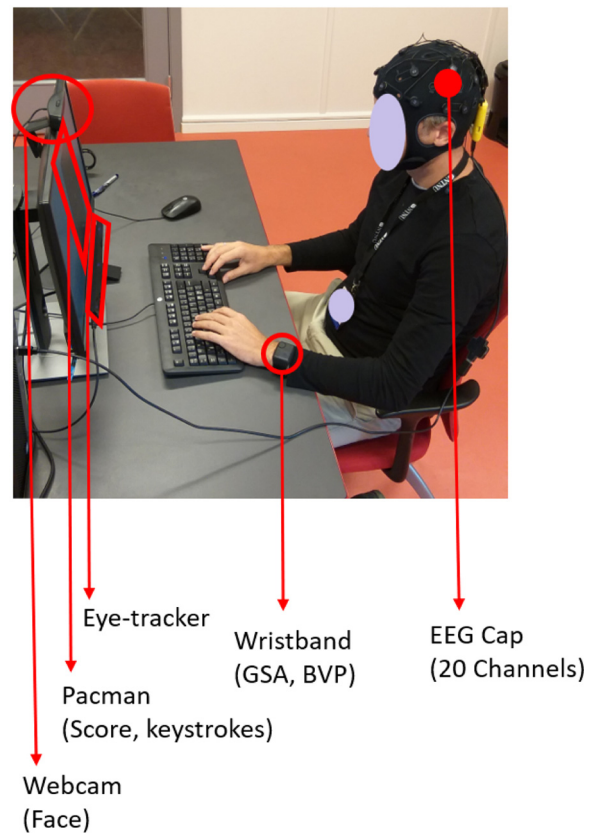

Fig. 2. Setup of the experiment, participants wear the wristband and EEG cap, calibrated to eye-tracking recording device and initiate the video and keystroke recording.

EEG and the wristband data streams were calibrated using the standard practices. The eye-tracker was calibrated using a 5-point calibration process; the EEG data was calibrated using the ENOBIO EOG correction mechanism. The researcher explained the mechanisms of the game and the respective keystrokes, double checked the data collection devices, and exited the room. The participant had approximately 40 min to master the game and achieve a score that was as high as possible.

3.4. Experimental design

The research design of our study is a single-group time series design (Ross and Morrison, 2004) with continuous (repeated) measurement of a group with the experimental treatment induced. Each participant played on average 16 game-sessions (SD = 7), until their allocated time ran out. Each game-session started with 3 lives and ended when the participant lost all the three lives. For each level in a game-session, the speed of the ghosts increased. Fig. 3 presents the protocol of our experiment. Each participant was shown a 5-s break before starting each session, then completed 2–3 min of game-play, and then had a 2–3 s reflection period while looking at their game score (Fig. 4).
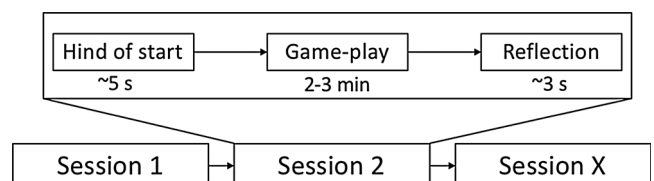

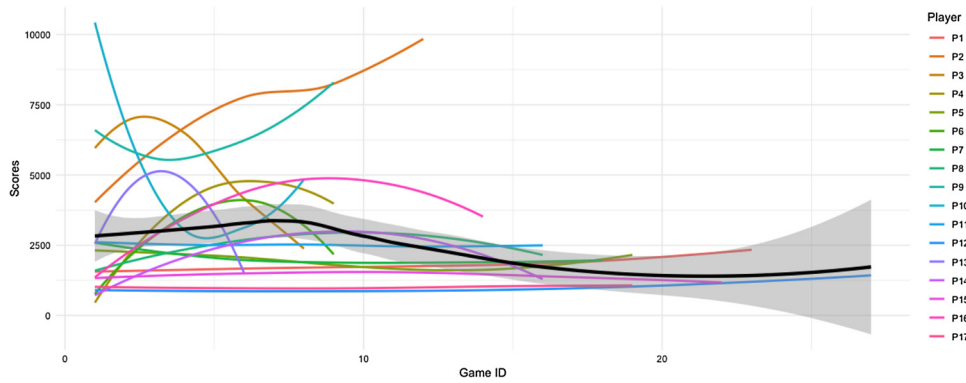**Fig. 3.** Protocol of the experiment.

Fig. 4. Score evolution or each user over-time.

3.5. Multimodal data collection setup

During the study we captured participants' achieved score for each game session. In addition, we collected sensor data from five different sources: keystrokes (representing click-stream data), eye-tracking, EEG, video, and wristband (with sensors for heart-rate, blood-pressure, temperature and electrodermal activity levels).

KeyStrokes – Participants' keystrokes were captured using Tobii's software. We used a full-size standard HP QWERTY keyboard (US English layout). The keys used were "N" to initiate a new game, and the arrow keys to move the Pac-Man. The raw keystroke data consist of key press and a time-stamp of when the key event occurred.

EEG – We recorded 20-channel EEG data organized in a standard 20 channel actiCAP layout following the international 10-20 system, as shown in Fig. 5. We built upon previous studies that utilize EEG headsets in detecting cognitive engagement in the learning domain

(Hassib et al., 2017a,b; Huang et al., 2014; Szafir and Mutlu, 2013). The raw EEG data was recorded at a 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 also applied an EOG filter to remove noise from blinks.

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 1920×1080 and the participants were 50–70 cm away from screen. Tobii's default algorithm was used to identify fixations and saccades (for details please see Olsen (2012)).

Video – Given the fact that we expected participants to exhibit minimal body and gesture information during the study, video recording focused on their face. We use a Logitech Web cam capturing video at 30 FPS. The webcam focus was zoomed 150% onto the faces of participants. The video resolution was 640×480.

Wristband – To record arousal data we use 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.

4. Results

From our collected data we want to identify how well these data associate with skill acquisition/development, and compare that to multimodal physiological models. We use game scores as a proxy for the level of skill acquired by the participants in each game session. In our analysis we explore relatively simple measurements such as face detection in the video, fixation/saccade measurements and simple frequency and auto correlation based features. We do not consider high-level features that require analysis of the stimulus itself, such as through object recognition, or using wavelet-transforms, or the definition of areas of interest.

Given the disparity of sampling rates of the different devices, we extract the features from each data stream separately. For the facial landmark detection we use the method described in (Kazemi and Sullivan, 2014) using Dlib and Python. The EEG and Keystroke features are computed using custom-written scripts in Matlab using the Signal Processing tool box and the eye-tracking features are computed using custom-written scripts in R. Then, the data from all the measurements is combined and analyzed using R. All the features are combined in two different ways. First, all the features from all the data streams were passed through LASSO for feature selection, the prediction results are marked by "All" in Fig. 7. Second, the features are selected for each data stream separately, and then passed through the prediction algorithm using different combinations of the data streams, which are explicitly marked in Fig. 7.

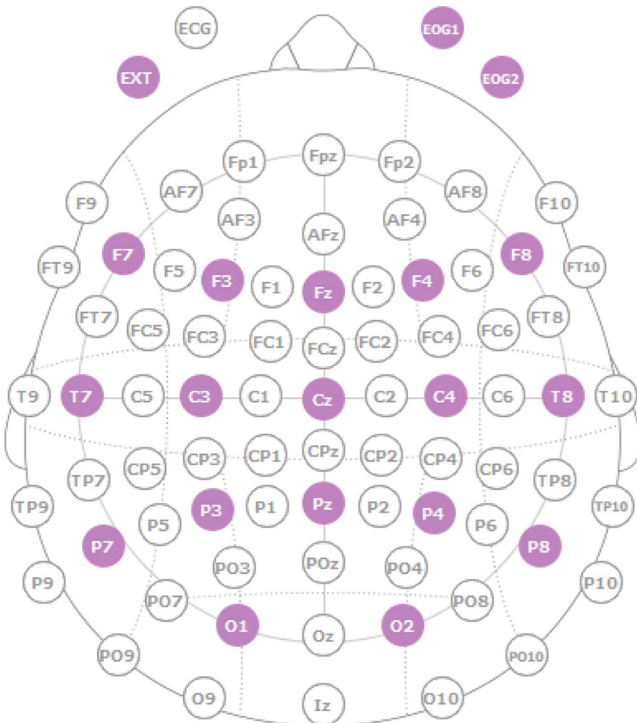**Fig. 5.** EEG electrode layout of 20 channels (EEG Layout for the experiment. The coloured ones are the electrodes being used. The white ones are those that the ENOBIO cap provides option for. This is the Standard electrode layout provided by the EEG capturing software. It is also known as the A-P (anterior–posterior) bipolar montage. This is also considered as Good "all-around" montage (Oostenveld and Praamstra, 2001).

4.1. Feature extraction

Given the limited sample in our study (17 participants, 251 game-sessions), the models are trained over the data of the sessions. We then extract features that have been used in literature to describe the respective data in previous studies. All the features were normalized with the average of the first 5 s of the data to remove the subjective bias from the data streams.

EEG – per channel: After obtaining data from the electrodes we normalize it between 0-1 and compute the first 10 Auto-correlation Coefficients[1] as proposed by Box et al. (2015) and further utilized in classification tasks by later work (Rahman et al., 2018; Wairagkar et al., 2015; Buscema et al., 2015). Auto-correlation coefficients describe the correlation between values of the same signal at different times, as a function of the time lags (time domain). To identify which frequency bands are more important, we compute the Fourier transform of the electrode signals and take the first 10 coefficients (first 10 dominant frequencies) (Sitnikova et al., 2009; Polat and Güneş, 2007; Guo et al., 2010) (Table 1).

Key Strokes: We initially collect raw keystroke data, consisting of key press and a time-stamp of when the key event occurred. We then derive our keystroke features from the timing information describing key press timing (i.e., keystroke dynamics). The only meaningful interaction through the keyboard was the arrow keys (up, down, left, right) to maneuver the Pacman and "N" to restart the game after loosing all the three lives. We create a time series based on all the key presses (considered in the same series) to perform feature extraction. Similarly to mouse behavior, keystroke dynamics seem promising for modeling user behavior (Epp et al., 2011; Vizer and Sears, 2017). To extract the features, we take the first 5 auto correlation coefficients, 6 linear predictive coefficients, energy of the signal, shape measures of the envelope of the signal (mean, variance, skew, kurtosis) and the first 10 Fourier transform coefficients. Table 2 provides the specific key strokes features.

Video-Face: Following the literature (Kazemi and Sullivan, 2014), to extract features from the videos we use as a guideline (Fig. 6). This allows us to extract the facial landmarks and then take the pair-wise distances between the points from the regions: right eye (15), left eye (15), right eyebrow (10), left eyebrow (10), nose (36), mouth (66) and jawline (136).

Gaze: Eye movement data provide the mean, variance, minimum, maximum and median of several parameters, such as pupil diameters, fixation details, saccade details, blink details, and event statistics. Table 3, provides an overview of the extracted features as well as the respective reference from the literature.

Wrist band: From the Empatica E4 wristband we extract the following features: mean, median, variance, skewness, maximum, minimum of (1) blood volume pressure, (2) EDA, (3) heart rate and (4) temperature. These are all the measures available from the recording device.

4.2. Feature selection

To select the most important features we employ the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). LASSO is an extension of ordinary least square (OLS) regression techniques fit for the cases where the number of examples are less than the length of the

---

[1] Auto-correlation coefficients depend on the sampling frequency, however, we used the same sampling frequency EEG for all our participants. Therefore, the dependency remains consistent for all the participants. Further, if someone uses the different sampling rate for the different participants, one way to have the minimal error in auto-correlation estimation is to sample the different EEG devices at the Nyquist frequency and then use the new samples to estimate the auto-correlation coefficients.

Table 1
EEG features.

Feature domain	Features extracted
Time domain	Auto correlation coefficients for the 10 previous lags
Frequency domain	Discrete Fourier Transform, first 10 coefficients

feature vector (Tibshirani, 1996). To find the best fitting curve for a set of data points, OLS tries to minimize the residual sum of squares (RSS) which is the difference between the actual values of the dependent variable (y) and the fitted values ($\hat{y}$). The formulation of the OLS is given as follows:

$$\hat{y} = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

The objective of the OLS regression is to minimize the difference between $\sum (\hat{y} - y)^2$ with the constraint that $\sum \beta_i^2 \leq s$. Where s is called the shrinkage factor.

LASSO on the other hand performs similar optimization with the slight difference in the constraint, which is now $\Sigma abs(\beta_i) \leq s$. While using LASSO, some of the β_i will be zero. Choosing s is like choosing the number of predictors in a regression model. Cross-validation can be used to estimate the best suited value for s.

Here, we use 5-fold cross validation to select the value of s. Our analysis seeks to identify how each of the 744 extracted features from the 5 different data-streams predicts motor-learning based on participants' game scores. We group our results along the five distinct categories of keystrokes, gaze, EEG, video, and wristband, and then apply the LASSO regression.[2] There are 65 selected features that are found to predict skill acquisition in our dataset (see Table 4).

4.3. Prediction results

To identify how the different data-streams can predict skill acquisition (measured from the game-score of the participants), we divide the whole data-set into training and testing sets, with data from one participant retained for testing. We perform a 17-fold cross validation (retaining one participant for testing each time) to remove the sampling bias from the training set. The normalized root mean squared error (NRMSE) values shown in Fig. 7 are the average values across all the cross validation folds for the testing sets.

To identify the contribution of each of the five data-streams in the prediction, we calculate and visualize the different NRMSEs obtained for every combination of data-streams (utilizing the respective features). To identify the impact of feature selection made by LASSO we calculate and visualize the different NRMSEs obtained when: (1) all the features are used (indicated by dots) and when (2) only selected by LASSO features are used (indicated by bars). Thus, we first use the Random Forest algorithm to predict the final score for every session, using all the 689 features (depicted with dots in Fig. 7), and then combine the different important features selected by LASSO (see Table 4). In addition, we use the Random Forest algorithm to predict the final score for every session (depicted with dots in Fig. 7. To summarize, Fig. 7 depicts the NRMSE prediction random forest: (1) by using only the selected features (the features selected by LASSO as presented in Table 4) – shown by bars and (2) by using all the extracted features (the 689 features extracted from the five data-streams) – shown by dots.

The results in Fig. 7 indicate that the combination of the selected features coming from eye-tracking, EEG and video data-streams provide the most accurate prediction of skill acquisition, while keystrokes alone provide the least accurate prediction. In addition, in every case the selected features (i.e., bars) provide more accurate prediction when

---

[2] We also execute a non-linear feature selection using Generalized Additive models (allowing for splines), this analysis produced similar results with LASSO.

Table 2
Key stroke features – all the arrow keys combined into one signal.

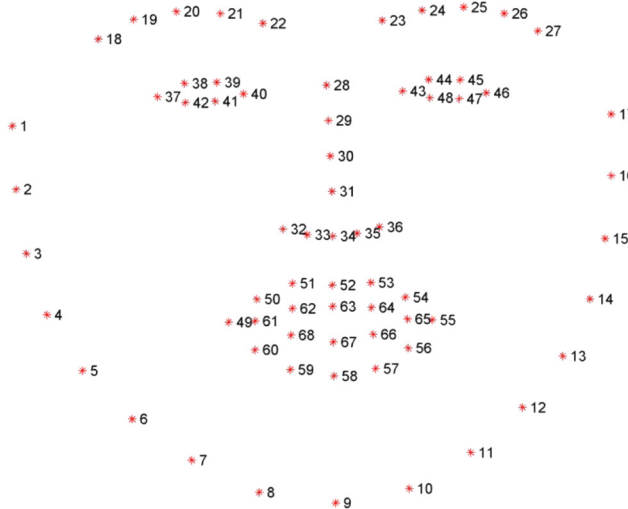| Feature domain	Features extracted
Time domain	Auto correlation coefficients for the 5 previous lags, envelope of the signal (mean, SD, skew, Kurtosis)
Frequency domain	Discrete Fourier transform, first 10 coefficients linear predictive coefficients (6), energy of the signal

Fig. 6. Face's landmark positions based on Kazemi and Sullivan (2014).

5. Discussion

Our results suggest that although different modalities can be a good proxy for skill acquisition (especially eye-tracking), fusing data from different modalities has the potential to further increase the accuracy of the prediction.

Our findings indicate that the modality that is used in the majority of studies (i.e., keystrokes) has the least accuracy (39% error, using all the features). For example, a recent literature review (Mangaroska and Giannakos, 2018) explains how only a few studies go beyond click-streams/keystrokes to inform learning design features, while another literature review on prediction in online courses found that only features related with user actions/preferences are used to build prediction models (Moreno-Marcos et al., 2019). Our results show that keystroke performance can be improved if the researchers perform feature selection (17% error), something that is in accordance with the literature (e.g., (Moreno-Marcos et al., 2019, 2018; Maldonado-Mahauad et al., 2018)). Yet, we find that fusing eye-tracking, EEG and video data-streams is the most accurate combination with an error of 6% (i.e., NRMSE 0.06). In other words, our results suggest that there is a sub-

Table 3
Eye-tracking features.

Eye movement parameters	Features extracted
Diameter	Pupil (Prieto et al., 2018) (mean, median, min, max, SD)
Fixation	Fixation duration (Reichle et al., 2009) (mean, median, min, max, SD)
	Fixation dispersion (Jaarsma et al., 2014) (mean, median, min, max, SD)
	Skewness of fixation duration histogram (Abernethy and Russell, 1987)
Saccade	Ratio of forward saccades to total saccades (Krischer and Zangemeister, 2007) (scanpath velocity)
	Ratio of global and local saccades (Zangemeister and Liman, 2007) (threshold on sac. vel.)
	Skewness of saccade velocity histogram (Liao et al., 2005)
	Saccade velocity (Russo et al., 2003) (mean, median, min, max, SD)
	Saccade length (mean, median, min, max, SD)
	Saccade amplitude (Phillips and Edelman, 2008) (mean, median, min, max, SD)
	Saccade duration (Vuori et al., 2004) (mean, median, min, max, SD)
Events	Num. Fixations, Num Saccades, fixation to saccade ratio

compared with the prediction coming from the all the 689 features (i.e., the whole dataset) and represented in Fig. 7 with dots.

The performance with modality fusion from eye-tracking, EEG and Video gives the optimal prediction with an error of 6% (i.e., NRMSE 0.06). To identify if the optimal prediction is significantly greater than any other combination, we used analysis of variances (ANOVA) to test it against the second best (i.e., eye-tracking) and found a significant difference ($F[1, 32] = 9.26$, $p < 0.05$). Thus our proposition that multi-modal data-streams have the capacity to give more accurate prediction of users' skills acquisition compared to traditional click-stream models has been confirmed. To go one step further we test if multimodal data that are *relatively easy-to-collect* (i.e., video, wrist-data, click-streams) have the capacity to give more accurate prediction compared to traditional click-stream models. Thus, we used an ANOVA to test the keystroke prediction against the one that combines keystrokes, video and wrist-data. Our results indicate a significant difference ($F[1, 32] = 5.01$, $p < 0.05$), confirming our proposition that even *relatively easy-to-collect* multimodal data have the capacity to obtain significantly greater predictions from traditional click-stream models.

stantial difference between what most studies in learning technology currently utilize (keystrokes), versus the expected potential of using multimodal data.

Taking into consideration both the technical difficulty and the high cost for the EEG and eye-tracking data collection, it is understandable why many studies in learning technology shy away from them. Nevertheless, researchers do not need to invest in expensive equipment and procedures to improve their models. Our results show that by simply combining data from the web camera and keystrokes can lead to an improved error rate (i.e., 15% error). Additionally, if a physiological wristband is available, then combining the web-camera with a wrist-band device like Empatica can reduce the error to 12%. While these approaches we recommend here may not achieve the optimal error rate of 6%, they are significantly convenient from a technical standpoint and do improve performance.

5.1. Contributions to knowledge

Understanding and supporting users' learning experience is still very

Table 4
Features selected by LASSO.

Data stream	Number of selected features	Top features selected by LASSO
Eye-tracking	9 (out of 42)	Saccade velocity (median, mean, max), Saccade length(SD, min), Saccade amplitude (max), Saccade duration (min, max, SD)
EEG	17 (out of 357)	First dominant frequency of all the channels θ-band (4–7 Hz) P_3, F_z, C_3 α-band (8–13 Hz) P_7, C_z, P_z, O_1, O_2, F_3 β-band (lower frequencies 13–20 Hz) P_4, P_8, T_8, F_8, C_4, F_4, T_7, F_7
Face video	18 (out of 290)	Jawline measures, left and right eye opening, mouth opening
Keystrokes	13 (out of 33)	Auto correlation coefficient (effect of past values), Linear predictive coefficient (coding shape of the signal), Linear frequency spectral coefficient (similar to LPC but immune to random peaks) Dominant frequencies (2, 5, 6, 7, 10)
Wristband	8 (out of 24)	Heart rate (mean, std. dev.), Electrodermal activity (mean, std. dev.), Body temperature (mean, std. dev.), Blood pressure (mean, std dev.)

sources of data. The benefit of any new multimodal data-stream lies in how it complements and works together with existing data-streams. It needs to be stated that multimodal data (including physiological) should not be seen as an attempt to replace, but rather to complement and supplement existing data-streams (e.g., clickstreams, self-reports). Our contribution is in accordance with similar works in other disciplines like ergonomics (Neuroergonomics) (Kramer and Parasuraman, 2007) and Information Systems (NeuroIS) (Dimoka et al., 2012, 2011), but introduces, for first time, five different data-streams and quantifies the predictive power compared to commonly used click-stream models. In addition, it investigates a very interesting area, that of learning systems, and exemplifies the great potential of multimodal- and physio-adaptive learning systems.

The value of multimodal data also comes from the fact that several latent variables cannot be measured objectively due to measurement biases, such as utility, and cognitive overload. Such variables can be measured reliably with multimodal data (Dimoka et al., 2012), enabling a new wave of research in the fields of systems' design (e.g., HCI and learning technologies). Nonetheless, differences between the various forms of multimodal data (e.g., physiological and computer logs or self-reports) should not necessarily imply that either approach is better. Instead it shows that there is a need for cross-validation when measuring complex constructs that are hard to capture accurately with a single data source. Differences between multimodal data may imply
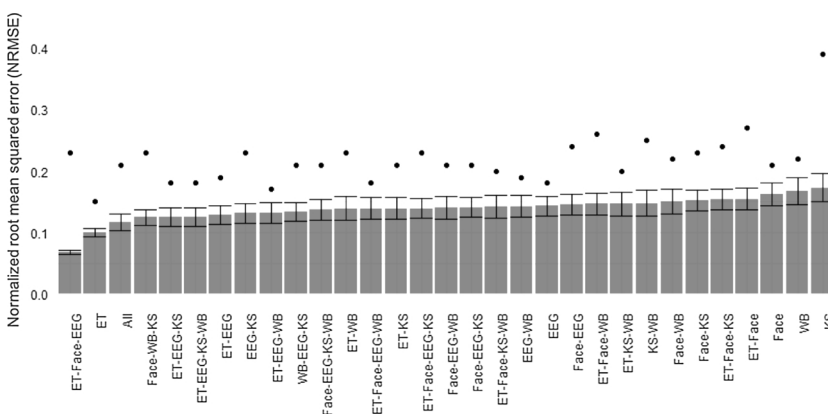

Fig. 7. Normalized root mean squared error (NRMSE) prediction from LASSO and random forest shown by bars (only with the selected features by LASSO as presented in Table 4), while the dots represent the prediction error when all the 689 features (i.e., the whole dataset) was used, the error-bars represent the 95% confidence interval.

limited, considering the wide range of multimodal data produced when a learner interacts with a system (Giannakos et al., 2018). Most of the work in the literature utilizes data coming from click-streams, keystrokes and self-reports (Moreno-Marcos et al., 2019; Maldonado-Mahauad et al., 2018). Predictive models are being used for anticipating learners' behavior and performance (Moreno-Marcos et al., 2018). Accurate prediction of behavior or achievements is very important in learning technology research (Gray and Perkins, 2019). Considering multimodal data in learning technologies (i.e., MMLA), from a predictive standpoint, this study clarifies that brain activity, facial reactions, arousal levels and gaze, can help us to better predict learner's behavior and achievements than the typical log data produced during the learner-system interaction. Since brain, gaze, facial features and arousal data are objective and do not suffer from measurement biases (Dimoka et al., 2012), our verified proposition in this paper is that they can be more reliable and accurate predictors (Bernheim, 2008). Since accuracy as well as early prediction are vital for contemporary learning systems (e.g., learner modeling, adaptive environments etc.), confirming and quantifying the benefits of multimodal data provides an evidence that learning technology research can rely on multimodal data to advance design systems that facilitate favorable behaviors and effectively enhance learning.

As noted in many of the examples offered in this paper, the value of multimodal data largely lies in combining physiological data with other

that either respondents are not willing or not able to self-report certain conditions (Dimoka et al., 2012), or simply that the human body simply cannot represent the richness of psychometric measures (Logothetis, 2008), or psychometric measures are not complex enough to capture the richness of the interaction with the system (e.g., cognitive load). Thus, besides increasing prediction accuracy, multimodal data also allows cross-validation and measurement of complex constructs that are hard (or even impossible) to capture with high accuracy using only a single data source.

Extending our work, the various multimodal data (e.g., brain data, gaze) might be able to better predict the success of most of the computer-based systems compared to self-reports. In addition, their high frequency and accuracy make them more appropriate for smaller sample sizes and short interventions (Dimoka et al., 2012). Another promising role of multimodal data is to inform debates that cannot be fully resolved with existing, most of the times single-source, data-sets. Many of the examples offered in this paper involve complex constructs that are still in their infancy or not fully democratized in the HCI and learning technology literature (e.g., cognitive states, cognitive load, mental state). In addition, the complexity that we can capture from multimodal data can be further examined through the configurational analysis prism. Recent works employ configurational analysis (e.g., fuzzy-set qualitative and comparative analysis – fsQCA) and combine self-reported data and data streams, leading to a deeper understanding

of the user (Papamitsiou et al., 2018; Pappas et al., 2019). Extending this work by fusing various multimodal data can have a significant impact in the way we currently view and understand the theories we use as well as the phenomena we study. Finally, multimodal data may also uncover new constructs (e.g., capturing temporary experience within a task) that have been ignored in the literature (e.g., they could not be adequately measured), thus furthering our understanding as well as the current theories.

5.2. Implications for practice

To gain further insight into the design of learning technologies, we consider the specific features that we find to be strongly associated with learning. Discussing these features from a technical standpoint can give rise to practical implications for the design of learning technologies.

First, our analysis of the eye-tracking data-stream revealed a number of feature categories as important learning predictors (skill acquisition). The first category includes features related to users' fatigue, anticipation and task difficulty (i.e., median, mean, max of saccade velocity) (McGregor and Stern, 1996; Bocca and Denise, 2006). This provides evidence that it is critical when we design learning technologies to monitor users' fatigue and relatively task difficulty and adapt the content accordingly. This is an unsurprising finding, as it is in accordance to the vast literature that proposes to keep the learner in a "flow experience" (Csikszentmihalyi, 2014), allowing them to engage with challenging, but not overwhelming experiences. Our results confirm this assertion, and show that using eye-tracking data it is possible to account for a large portion in the variation of learning.

The second important eye-tracking feature set relates to user's ability to see specific information (i.e., SD and min of saccade length). The implication of this finding is that designing environments that allow users to easily identify the details and avoid long saccades (e.g., big dashboards) can improve learning (Russo et al., 2003; Phillips and Edelman, 2008). Another important feature set relates to user's high cognitive load (i.e., max saccade amplitude (May et al., 1990; Goldberg et al., 2002)) and high task difficulty (i.e., min, max, SD saccade duration (Vuori et al., 2004; Bestelmeyer et al., 2006)). Given that cognitive load and task difficulty indicate how much "space" in working memory is currently being used (Sweller, 2011), the observed strong association with learning was expected. This verifies previous studies, but also indicates the potential of integrating the germane cognitive load principles into HCI (Hollender et al., 2010).

In the EEG data-stream we identify the Fourier coefficient of the most dominant frequency in each channel to be the most predictive. We further analyze the signal from every EEG channel to find the most dominant frequency band present, resulting to the identification of the theta band (4–7 Hz, 3 channels), alpha band (8–18 Hz, 6 channels) and lower beta band (13–20 Hz, 8 channels). These frequency bands are strongly associated with depicting attention, emotion, load on the memory, and cognitive load respectively. For example, the presence of theta band reflects load on working and/or episodic memory (Kiiski et al., 2017; Dai et al., 2017; Klimesch et al., 1994), the alpha band indicates attention (Klimesch et al., 1998; Shaw, 1996) and arousal (alertness, (Gazzaniga, 2012)), both alpha and theta bands reflect cognitive load while solving problems (Kiiski et al., 2017; Dai et al., 2017; Klimesch et al., 1994), and finally, the beta band has been shown to reflect emotional process (Ray and Cole, 1985) and inattention (Roh et al., 2016; Loo et al., 2018). Since our findings show that these bands are strongly associated with learning, we argue that these processes should be considered when designing the User Interface or the overall User Experience, to improve learning outcomes.

In the face-video data-stream we identified important features that relate to jawline measures, mouth opening, and eyes opening. These features relate to affective states. In affective computing, and in particular affective learning technologies (Kapoor and Picard, 2005; Whitehill et al., 2014), these states have been used to create affective

instructional strategies (Bosch et al., 2015). Such strategies can offer valuable information in advancing learning technologies (Bosch et al., 2015; Wiggins et al., 2015; Rana and Dwivedi, 2017), such as providing information that a user is engaged or delighted with the current task.

In the wristband data-stream, the important features are mean and SD for heart rate, electrodermal activity, body temperature and blood pressure. These are features associated with users' affective states (Slovák et al., 2014; Hassib et al., 2016). Given that the core idea is to utilize multimodal data to enhance the learning experience, one promising strategy is to identify and minimize negative affective states (e.g. frustration, confusion, boredom) through a learning technology that has the capacity to utilize learner's affective state. Our findings quantify how much benefit we can expect to gain by incorporating such measures in the design of learning systems.

Finally, in the more traditional keystroke data-stream, the important features relate to stability/predictability of a user's input (i.e., auto correlation coefficient), input behavior/pattern (i.e., linear predictive coefficient), duration of the patterns (i.e., mean of envelope shape) and the highest point of activity (i.e., the most Dominant Frequency). Such keystroke dynamics are associated with users' cognitive and affective states, like mood and emotions (Epp et al., 2011), and cognitive and physical stress (Vizer and Sears, 2017; Vizer, 2013). As a baseline performance improvement, our work shows that if only keystroke data is available, then it is prudent to perform feature selection on the data, rather than using all available features, since that can almost half the achieved error rate without incorporating additional modalities.

Overall, our results point to the triangulation of insights coming from different modalities as a means to achieve better (but also faster) identification of learners' cognitive-affective states. If an LCI scenario has the capability to collect multimodal data from learners' during their interaction with the system, it is possible to obtain a multifaceted understanding of whether the learner is exerting mental effort and/or needs time or help. Of course, once a remedial action is implemented (e.g., provide a hint, reduce the task difficulty), the same measures may offer an insight into whether this was effective at all and then inform the machine back.

5.3. Limitations

Our findings speak in favor of collecting eye-tracking and EEG data to predict learner performance (in our case skill development), but are subject to certain limitations. Our study participants – undergraduate and graduate students – represent an appropriate sample for a study on learning technologies (covering the higher education population), but are less representative of the population of K-12 schools (e.g., classroom learning), lifelong learning and learning in workplace who rely heavily on learning systems as well. Moreover, the study was performed in a controlled environment, which may have induced demand characteristics that affect the ecology of the study (e.g., participants' performance and behavior). Specifically, participants were aware of the multimodal data collection since they signed a detailed consent form, which may have led to increased apprehension and desire "to perform". Nonetheless, the population represents a large part of the learning technology end-users and the data collections are of very high quality and accuracy utilizing state of the art equipment.

In our study, we utilize a time-tested game that has very shallow learning curve and has been used to measure basic skills development in the past (Nicolson and Fawcett, 2000). The generalizability of our findings is constrained by the significant variation of the interactions someone has with learning technologies (e.g., reading/watching materials, solving problems etc.), different types of learning (e.g., cognitive learning instead of skill acquisition), as well as the way learning is captured (e.g., knowledge tests or other assessment mechanisms instead of the score of the game). Although researchers like Grissmer et al. (2010) have shown that, indeed, skill development is a strong predictor

of cognitive learning performance (by analyzing data from six different data-sets), we recognize that considering additional learning tasks and stages of learning will offer a holistic understanding of the role of multimodal data. However, this study was a first of its kind, and as a springboard for future research studies it employs a stimulus widely used in learning (i.e., a game) and uses the score of the game as a proxy for the skill acquisition.

Finally, we captured specific multimodal data, namely keystrokes, EEG, eye-tracking, video and wristband data. In other more complex learning tasks (e.g., with embodied interaction, more complex input devices) it would have been possible to include additional data-streams and features (e.g., gestures (Karambakhsh et al., 2018)). Moreover, additional EEG channels or post-hoc features (e.g., areas of interest) may have offered additional insights. Thus, although we try to include as many data-streams as possible, we understand that other multimodal data sources may also play an important role. However, our selection includes a rich set of data-streams that are common to typical interactions with contemporary learning systems.

6. Conclusion and ongoing work

Overall, our work shows that capturing multimodal data can help us increase the prediction accuracy of users' learning performance in learner–computer interaction (LCI). In addition, the study shows that the most commonly used data-stream (i.e., keystrokes) is the poorest proxy of our learning performance. Thus, leveraging advances in contemporary learning environments and physiological sensing (wearable, EEG etc.), we provide evidence that multimodal data can be a viable method to accurately track users' states during learning, thereby providing unique possibilities of closing the loop between the learning technology and the learner. Therefore, the incorporation of multimodal data enables HCI and learning technology researchers to examine unscripted, complex tasks in more holistic and accurate ways.

The contribution of this paper is threefold: (1) we conduct a controlled study that collects five different data-streams during a basic LCI task and select the features of those data-streams that were found to be more important for skill development of the user, (2) we quantify the capacity of various data-streams to give more accurate prediction of users' acquired skill and (3) we discuss how these findings (i.e., selected features, data-streams prioritization) can be used for the design of future learning technologies as well as for advancing the research design of future experiments in the area of learning technologies.

For our future work, we are now beginning to collect multimodal data from different types of LCI (e.g., MOOCs, assessment) utilizing different input devices (e.g., combination of mouse-keyboard) as well as embodied interaction. In addition, we intend to investigate whether a plausible association exists between different user-groups (e.g., age, skills, gender) or stimulus used in LCI and the produced multimodal data.

After collecting multimodal data-sets from different LCI contexts, we will be able to identify the data-streams and features of those streams that are strongly associated with different learning tasks. Such a mapping will allow us to select a set of data-steams and features that can be employed in out-of-the-lab context, but also give very accurate results. Thus, we want to enable "In-the-Wild" studies, attain high ecology as well as induce them to contemporary research in learning technology (e.g., personalized learning, learner models etc.).

Acknowledgments

References

Abdelrahman, Y., Velloso, E., Dingler, T., Schmidt, A., & Vetere, F. (2017 September). Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(3), https://doi.org/10.1145/3130898 33:1–33:20.

Abernethy, B., & Russell, D. G. (1987). Expert-novice differences in an applied selective attention task. *Journal of Sport Psychology, 9*(4), 326–345.

Ashenafi, M. M., Ronchetti, M., & Riccardi, G. (2016). Predicting student progress from peer-assessment data. *Proceedings of the 9th international conference on educational data mining,* 270–275.

Bernheim, B. D. (2008). Neuroeconomics: A sober-but hopeful-appraisal. *NBER Working Paper Series 13954*.

Bestelmeyer, P. E., Tatler, B. W., Phillips, L. H., Fraser, G., Benson, P. J., & Clair, D. S. (2006). Global visual scanning abnormalities in schizophrenia and bipolar disorder. *Schizophrenia Research, 87*(1-3), 212–222.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3*(2), 220–238.

Bocca, M.-L., & Denise, P. (2006). Total sleep deprivation effect on disengagement of spatial attention as assessed by saccadic eye movements. *Clinical Neurophysiology, 117*(4), 894–899.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., ... Ventura, M., Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th international conference on intelligent user interfaces,* 379–388.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control.* John Wiley & Sons.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602–607.

Bulling, A., Ward, J. A., Gellersen, H., & Troster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(4), 741–753.

Buscema, M., Vernieri, F., Massini, G., Scrascia, F., Breda, M., Rossini, P. M., & Grossi, E. (2015). An improved i-fast system for the diagnosis of Alzheimer's disease from unprocessed electroencephalograms by using robust invariant features. *Artificial Intelligence in Medicine, 64*(1), 59–74.

Chang, V. (2016). Review and discussion: E-learning for academia and industry. *International Journal of Information Management, 36*(3), 476–485.

Csikszentmihalyi, M. (2014). *Learning, "flow," and happiness. Applications of flow in human development and education.* Springer153–172.

Dai, Z., De Souza, J., Lim, J., Ho, P. M., Chen, Y., Li, J., ... Sun, Y. (2017). EEG cortical connectivity analysis of working memory reveals topological reorganization in theta and alpha bands. *Frontiers in Human Neuroscience, 11,* 237.

Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., & Specht, M. (2017). Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. *Proceedings of the seventh international learning analytics & knowledge conference,* 188–197.

Dimoka, A., Davis, F. D., Gupta, A., Pavlou, P. A., Banker, R. D., Dennis, A. R., ... Weber, B. (2012). On the use of neurophysiological tools in IS research: Developing a research agenda for neurois. *MIS Quarterly,* 679–702.

Dimoka, A., Pavlou, P. A., & Davis, F. D. (2011). Research commentary—Neurois: The potential of cognitive neuroscience for information systems research. *Information Systems Research, 22*(4), 687–702.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR), 47*(3), 43.

Duchowski, A. T. (2007). *Eye tracking methodology. Theory and practice*328.

Echeverria, V., Martinez-Maldonado, R., Power, T., Hayes, C., & Shum, S. B. (2018). Where is the nurse? Towards automatically visualising meaningful team movement in healthcare education. *International conference on artificial intelligence in education,* 74–78.

EdX (2018). *Edx research guide release.* Retrieved from https://media.readthedocs.org/pdf/devdata/latest/devdata.pdf.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer, 49*(4), 61–69.

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. *Proceedings of the Sigchi conference on human factors in computing systems,* 715–724.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (2018). *The Cambridge handbook of expertise and expert performance.* Cambridge University Press.

Exler, A., Schankin, A., Klebsattel, C., & Beigl, M. (2016). A wearable system for mood assessment considering smartphone features and data from mobile ECGS. *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct,* 1153–1161.

Feit, A. M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., & Morris, M. R. (2017). Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. *Proceedings of the 2017 Chi conference on human factors in computing systems,* 1118–1130.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gazzaniga, M. S. (2012). *Handbook of psychobiology.* Elsevier.

Giannakos, M., Sharma, K., Martinez-Maldonado, R., Dillenbourg, P., & Rogers, Y. (2018).

Learner–computer interaction. Proceedings of the 10th nordic conference on human–computer interaction. New York, NY, USA: ACM https://doi.org/10.1145/3240167.3240259.

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *The International Review of Research in Open and Distributed Learning, 16*(1).

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002). Eye tracking in web search tasks: Design implications. *Proceedings of the 2002 symposium on eye tracking research and applications,* 51–58.

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education, 131,* 22–32.

Grissmer, D., Grimm, K. J., Aiyer, S. M., Murrah, W. M., & Steele, J. S. (2010). Fine motor skills and early comprehension of the world: Two new school readiness indicators. *Developmental Psychology, 46*(5), 1008.

Guo, L., Rivero, D., Dorado, J., Rabunal, J. R., & Pazos, A. (2010). Automatic epileptic seizure detection in EEGS based on line length feature and artificial neural networks. *Journal of Neuroscience Methods, 191*(1), 101–109.

Hassib, M., Khamis, M., Friedl, S., Schneegass, S., & Alt, F. (2017a]). Brainatwork: Logging cognitive engagement and tasks in the workplace using electroencephalography. *Proceedings of the 16th international conference on mobile and ubiquitous multimedia,* 305–310.

Hassib, M., Khamis, M., Schneegass, S., Shirazi, A. S., & Alt, F. (2016). Investigating user needs for bio-sensing and affective wearables. *Proceedings of the 2016 Chi conference extended abstracts on human factors in computing systems,* 1415–1422.

Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., & Alt, F. (2017b]). Engagemeter: A system for implicit audience engagement sensing using electroencephalography. *Proceedings of the 2017 Chi conference on human factors in computing systems,* 5114–5119.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior, 26*(6), 1278–1288.

Huang, J., Yu, C., Wang, Y., Zhao, Y., Liu, S., Mo, C., ... Shi, Y. (2014). Focus: Enhancing children's engagement in reading by using contextual BCI training sessions. *Proceedings of the 32nd annual ACM conference on human factors in computing systems,* 1905–1908.

Jaarsma, T., Jarodzka, H., Nap, M., van Merrienboer, J. J., & Boshuizen, H. P. (2014). Expertise under the microscope: Processing histopathological slides. *Medical Education, 48*(3), 292–300.

Kanjo, E., Al-Husain, L., & Chamberlain, A. (2015). Emotions in context: Examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing, 19*(7), 1197–1212.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. *Proceedings of the 13th annual ACM international conference on multimedia,* 677–682.

Karambakhsh, A., Kamel, A., Sheng, B., Li, P., Yang, P., & Feng, D. D. (2018). Deep gesture interaction for augmented anatomy learning. *International Journal of Information Management.*

Katerina Mangaroska, B. V., & Giannakos, M. (2019). Cross-platform analytics: A step towards personalization and adaptation in education. *Proceedings of the 9th international learning analytics and knowledge conference (LAK19),* 243–247.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1867–1874.

Kiiski, H., Bennett, M., Kelly, C., Grogan, K., Rai, L., Sweeney, A., ... Whelan, R. (2017). EEG theta/beta ratio correlates positively with inattention and memory problems in ADHD and controls. *European Neuropsychopharmacology, 27,* S1102–S1103.

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., & Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neuroscience Letters, 244*(2), 73–76.

Klimesch, W., Schimke, H., & Schwaiger, J. (1994). Episodic and semantic memory: An analysis in the EEG theta and alpha band. *Electroencephalography and Clinical Neurophysiology, 91*(6), 428–441.

Krakauer, J. W., & Mazzoni, P. (2011). Human sensorimotor learning: Adaptation, skill, and beyond. *Current Opinion in Neurobiology, 21*(4), 636–644.

Kramer, A. F., & Parasuraman, R. (2007). *Neuroergonomics: Applications of neuroscience to human factors.*

Krischer, C., & Zangemeister, W. H. (2007). Scanpaths in reading and picture viewing: Computer-assisted optimization of display conditions. *Computers in Biology and Medicine, 37*(7), 947–956.

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE, 103*(9), 1449–1477.

Lai, N. K., Ang, T. F., Por, L. Y., & Liew, C. S. (2018). Learning through intuitive interface: A case study on preschool learning. *Computers & Education, 126,* 443–458.

Lee, D. H., & Anderson, A. K. (2017). Reading what the mind thinks from how the eye sees. *Psychological Science, 28*(4), 494–503.

Liao, K., Kumar, A. N., Han, Y. H., Grammer, V. A., Gedeon, B. T., & Leigh, R. J. (2005). Comparison of velocity waveforms of eye and head saccades. *Annals of the New York Academy of Sciences, 1039*(1), 477–479.

Logothetis, N. K. (2008). What we can do and what we cannot do with FMRI. *Nature, 453*(7197), 869.

Loo, S. K., McGough, J. J., McCracken, J. T., & Smalley, S. L. (2018). Parsing heterogeneity in attention-deficit hyperactivity disorder using EEG-based subgroups. *Journal of Child Psychology and Psychiatry, 59*(3), 223–231.

Maior, H. A., Wilson, M. L., & Sharples, S. (2018). Workload alerts-using physiological measures of mental workload to provide feedback during tasks. *ACM Transactions on Computer-Human Interaction (TOCHI), 25*(2), 9.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Mu noz-Merino, P. J., & Delgado-Kloos, C. (2018). Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning. *European conference on technology enhanced learning,* 355–369.

Mangaroska, K., & Giannakos, M. (2018). Comlearning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies.* https://doi.org/10.1109/TLT.2018.2868673.

Manh aes, L. M. B., da Cruz, S. M. S., & Zimbr ao, G. (2014). Wave: An architecture for predicting dropout in undergraduate courses using EDM. *Proceedings of the 29th annual ACM symposium on applied computing,* 243–247.

Martinez-Maldonado, R., Power, T., Hayes, C., Abdiprano, A., Vo, T., Axisa, C., & Buckingham Shum, S. (2017). Analytics meet patient manikins: Challenges in an authentic small-group healthcare simulation classroom. *Proceedings of the seventh international learning analytics & knowledge conference,* 90–94.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica, 75*(1), 75–89.

McGregor, D. K., & Stern, J. A. (1996). Time on task and blink effects on saccade duration. *Ergonomics, 39*(4), 649–660.

Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on: Detecting cognitive load using EEG during learning. *Proceedings of the seventh international learning analytics & knowledge conference,* 80–89.

Moreno-Marcos, P. M., Alario-Hoyos, C., Mu noz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCS: A review and future research directions. *IEEE Transactions on Learning Technologies.*

Moreno-Marcos, P. M., Mu noz-Merino, P. J., Alario-Hoyos, C., Estévez-Ayres, I., & Delgado Kloos, C. (2018). Analysing the predictive power for anticipating assignment grades in a massive open online course. *Behaviour & Information Technology,* 1–16.

Nicolson, R. I., & Fawcett, A. J. (2000). Long-term learning in dyslexic children. *European Journal of Cognitive Psychology, 12*(3), 357–393.

Niforatos, E., Vourvopoulos, A., & Langheinrich, M. (2017). Amplifying human cognition: Bridging the cognitive gap between human and machine. *Proceedings of the 2017 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 ACM international symposium on wearable computers,* 673–680.

Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B., & Castells, J. (2013). Expertise estimation based on simple multimodal features. *Proceedings of the 15th ACM on international conference on multimodal interaction,* 583–590.

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The rap system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. *Proceedings of the 8th international conference on learning analytics and knowledge,* 360–364.

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data? *Journal of Learning Analytics, 3*(2), 213–219.

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

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology, 112*(4), 713–719.

Panigrahi, R., Srivastava, P. R., & Sharma, D. (2018). Online learning: Adoption, continuance, and learning outcome—A review of literature. *International Journal of Information Management, 43,* 1–14.

Pantazos, K., & Vatrapu, R. (2016). Enhancing the professional vision of teachers: A physiological study of teaching analytics dashboards of students' repertory grid exercises in business education. *2016 49th Hawaii international conference on System sciences (HICSS),* 41–50.

Papamitsiou, Z., Economides, A. A., Pappas, I. O., & Giannakos, M. N. (2018). Explaining learning performance using response-time, self-regulation and satisfaction from content: An FSQCA approach. *Proceedings of the 8th international conference on learning analytics and knowledge,* 181–190.

Pappas, I. O., Giannakos, M. N., & Sampson, D. G. (2019 March). Fuzzy set analysis as a means to understand users of 21st-century learning systems: The case of mobile learning and reflections on learning analytics research. *Computers in Human Behavior, 92,* 646–659. https://doi.org/10.1016/j.chb.2017.10.010.

Pardo, A., Han, F., & Ellis, R. A. (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), 82–92.

Phillips, M. H., & Edelman, J. A. (2008). The dependence of visual scanning performance on search direction and difficulty. *Vision Research, 48*(21), 2184–2192.

Polat, K., & Güneş, S. (2007). Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation, 187*(2), 1017–1026.

Prieto, L. P., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M. J., & Dillenbourg, P. (2018). Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning, 34*(2), 193–203.

Raca, M., Kidzinski, L., & Dillenbourg, P. (2015). Translating head motion into attention-towards processing of student's body-language. *In Proceedings of the 8th international conference on educational data mining* (pp. 320–326).

Rahman, M. M., Chowdhury, M., & Fattah, S. A. (2018). An efficient scheme for mental task classification utilizing reflection coefficients obtained from autocorrelation function of EEG signal. *Brain Informatics, 5*(1), 1.

Rana, N. P., & Dwivedi, Y. K. (2016). Using clickers in a large business class: Examining use behavior and satisfaction. *Journal of Marketing Education, 38*(1), 47–64.

Rana, N. P., & Dwivedi, Y. K. (2017). Can clicking promote learning? Measuring student learning performance using clickers in the undergraduate information systems class. *Journal of International Education in Business, 10*(2), 201–215.

Rana, N. P., & Dwivedi, Y. K. (2018). An empirical examination of antecedents determining students' usage of clickers in a digital marketing module. *International Journal of Business Information Systems, 27*(1), 86–104.

Rana, N. P., Dwivedi, Y. K., & Al-Khowaiter, W. A. (2016). A review of literature on the use of clickers in the business and management discipline. *The International Journal of Management Education, 14*(2), 74–91.

Ray, W. J., & Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science, 228*(4700), 750–752.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1–21.

Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting performance on MOOC assessments using multi-regression models. *Proceedings of the 9th international conference on educational data mining,* 484–489.

Roh, S.-C., Park, E.-J., Shim, M., & Lee, S.-H. (2016). EEG beta and low gamma power correlates with inattention in patients with major depressive disorder. *Journal of Affective Disorders, 204*, 124–130.

Ross, S. M., & Morrison, G. R. (2004). Experimental research methods. *Handbook of research on educational communications and technology, 2*, 1021–1043.

Russo, M., Thomas, M., Thorne, D., Sing, H., Redmond, D., Rowland, L., ... Balkin, T. (2003). Oculomotor impairment during chronic partial sleep deprivation. *Clinical Neurophysiology, 114*(4), 723–736.

Scheidt, S., & Chung, Q. (2018). Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation. *International Journal of Information Management*.

Shaw, J. C. (1996). Intention as a component of the alpha-rhythm response to mental activity. *International Journal of Psychophysiology, 24*(1-2), 7–23.

Sitnikova, E., Hramov, A. E., Koronovsky, A. A., & van Luijtelaar, G. (2009). Sleep spindles and spike-wave discharges in EEG: Their generic features, similarities and distinctions disclosed with fourier transform and continuous wavelet analysis. *Journal of Neuroscience Methods, 180*(2), 304–316.

Slovák, P., Tennent, P., Reeves, S., & Fitzpatrick, G. (2014). Exploring skin conductance synchronisation in everyday interactions. *Proceedings of the 8th Nordic conference on human–computer interaction: Fun, fast, foundational,* 511–520.

Sweller, J. (2011). *Cognitive load theory. Psychology of learning and motivation, Vol. 55*, Elsevier37–76.

Szafir, D., & Mutlu, B. (2013). Artful: Adaptive review technology for flipped learning. *Proceedings of the Sigchi conference on human factors in computing systems,* 1001–1010.

Teplan, M., et al. (2002). Fundamentals of EEG measurement. *Measurement Science Review, 2*(2), 1–11.

Thon, B. (2015). Cognition and motor skill learning. *Annals of Physical and Rehabilitation Medicine, 58*, e25.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological),* 267–288.

Tursunbayeva, A., Di Lauro, S., & Pagliari, C. (2018). People analytics—A scoping review of conceptual boundaries and value propositions. *International Journal of Information Management, 43*, 224–247.

Vizer, L. M. (2013). Different strokes for different folks: individual stress response as manifested in typed text. *Chi'13 extended abstracts on human factors in computing systems,* 2773–2778.

Vizer, L. M., & Sears, A. (2017). Efficacy of personalized models in discriminating high cognitive demand conditions using text-based interactions. *International Journal of Human-Computer Studies, 104*, 80–96.

Vuori, T., Olkkonen, M., Pölönen, M., Siren, A., & Häkkinen, J. (2004). Can eye movements be quantitatively applied to image quality studies? *Proceedings of the third Nordic conference on human–computer interaction,* 335–338.

Wairagkar, M., Hayashi, Y., & Nasuto, S. (2015). Movement intention detection from autocorrelation of EEG for BCI. *International conference on brain informatics and health,* 212–221.

Wang, G., Zhang, X., Tang, S., Zheng, H., & Zhao, B. Y. (2016). Unsupervised clickstream clustering for user behavior analysis. *Proceedings of the 2016 Chi conference on human factors in computing systems,* 225–236.

Watson, C., Li, F. W., & Godwin, J. L. (2013). Predicting performance in an introductory programming course by logging and analyzing student programming behavior. *2013 IEEE 13th international conference on advanced learning technologies (ICALT),* 319–323.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing, 5*(1), 86–98.

Wiggins, J. B., Boyer, K. E., Baikadi, A., Ezen-Can, A., Grafsgaard, J. F., Ha, E. Y., ... Wiebe, E. N. (2015). Javatutor: An intelligent tutoring system that adapts to cognitive and affective states during computer programming. *Proceedings of the 46th ACM technical symposium on computer science education,* 599.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience, 12*(12), 739.

Yousafzai, A., Chang, V., Gani, A., & Noor, R. M. (2016). Multimedia augmented m-learning: Issues, trends and open challenges. *International Journal of Information Management, 36*(5), 784–792.

Zangemeister, W. H., & Liman, T. (2007). Foveal versus parafoveal scanpaths of visual imagery in virtual hemianopic subjects. *Computers in Biology and Medicine, 37*(7), 975–982.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76–86.

Zhao, C., Yang, J., Liang, J., & Li, C. (2016). Discover learning behavior patterns to predict certification. *2016 11th international conference on Computer science and education (ICCSE),* 69–73.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., & Cichocki, A. (2018). Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics, 99*, 1–13.