International Journal of Human-Computer Studies 136 (2020) 102384

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Fitbit for learning: Towards capturing the learning experience using wearable sensing

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

[a] Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] The University of Melbourne, Australia
[c] University of Agder (UiA), Kristiansand, Norway

ARTICLE INFO

Keywords:
Wearables
Learning experience
Sensing
Machine learning
Tracking learning

ABSTRACT

The assessment of learning during class activities mostly relies on standardized questionnaires to evaluate the efficacy of the learning design elements. However, standardized questionnaires pose additional strain on students, do not provide "temporal" information during the learning experience, require considerable effort and language competence, and sometimes are not appropriate. To overcome these challenges, we propose using wearable devices, which allow for continuous and unobtrusive monitoring of physiological parameters during learning. In this paper we set out to quantify how well we can infer students' learning experience from wrist-worn devices capturing physiological data. We collected data from 31 students in 93 class sessions (3 class sessions per student), and our analysis shows that wrist data can predict the learning experience with 11% error. We also show that 6.25 min (SD = 3.1 min) of data are needed to achieve a reliable estimate (i.e., 13.8% error). Our work highlights the benefits and limitations of utilizing wearable devices to assess learning experiences. Our findings help shape the future of quantified-self technologies in learning by pointing out the substantial benefits of physiological sensing for self-monitoring, evaluation, and metacognitive reflection in learning.

1. Introduction

Positive learning experience provides learners with chunks of time that intentionally propel them towards their learning goals (Schmidt et al., 2019). One's learning experience changes over time and is dependent on interventions, social interactions and changing contexts (Fredricks et al., 2016). Having the ability to assess students' experience within different contexts and phases of learning, utilizing self-monitoring, self-evaluation, and metacognitive reflection, can help us to improve the contemporary design of teaching and learning.

To assess students' learning experience, several questionnaire instruments have been developed and widely used in the past (Kay and Knaack, 2009; Henrie et al., 2015). Such instruments have been found useful in informing students, instructors, and educational institutions regarding important teaching decisions (Kuh, 2001). Despite the value of these questionnaire instruments, however, such an approach is not always ideal for assessing students' experience during learning (Aslan et al., 2017). For instance, for questionnaires to be valid and reliable, students need to invest time and effort, and it is not possible to apply these questionnaires multiple times to assess all courses of a semester. In addition, even if this were possible, the questionnaire would assess the overall experience and it would not be possible to draw accurate conclusions about specific learning designs, aspects, and periods of the course. In addition, questionnaires can be inappropriate for some students (e.g., those in primary education, or those with special needs, who may not fully grasp the questions). We argue that if physiological data can accurately infer the responses from standardized questionnaires while we are collecting ecologically valid responses, then learning experience can be measured continuously and unobtrusively, without having to disrupt learning or require effort from students.

Wearable devices that enable physiological sensing are now widely available and affordable, providing the capacity to obtain and store everyday data about one's routine activities (Choe et al., 2014). At the same time, it has become possible, through these devices, to monitor more subtle phenomena, such as the quality of social interactions, students' mental health, and learning engagement (Hernandez et al., 2014; Wang et al., 2018). The quantified-self movement has shown its potential to utilize authentic and granular activity data in order to inform users about their lifestyle and fitness (Lee et al., 2016), with the

* Corresponding author.
E-mail address: michailg@ntnu.no (M.N. Giannakos).

https://doi.org/10.1016/j.ijhcs.2019.102384
Received 13 November 2018; Received in revised form 29 November 2019; Accepted 10 December 2019
Available online 11 December 2019

aim of involving the user in self-monitoring and self-reflection processes to regulate different aspects of their life and behavior (Ruiz et al., 2016). This has been an intriguing sociotechnical development of the last decade, and allows us to leverage different technologies and sensors (e.g., wristbands, cameras) to monitor and utilize human physiological parameters. Despite this great potential for monitoring learning, however, the direction remains rather underexplored (Henrie et al., 2015).

To investigate the potential of wearable sensing to monitor learning experience, we form the following research questions:

- Can we accurately estimate students' learning experience from sensing data obtained from wrist-worn devices?
- How much time do we need to achieve an accurate estimation of students' learning experience?

To tackle the aforementioned research questions, we conduct a study in which we use wrist-worn devices to capture physiological data from students during class activities. The data are augmented by standardized questionnaires completed at the end of each activity. We then apply machine learning techniques to infer learning experience from the physiological data (calculating root mean square error [RMSE]), and identify the minimum time window needed to make a reliable prediction. By investigating the technology's feasibility and identifying the minimum data required, we provide a path towards the design of technology that overcomes the disadvantages of traditional standardized instruments of learning experience. If information about students' experience during lectures can be made available to teachers, they too can self-reflect on their teaching performance and design effective methods for student-centered learning design. However, to enable the creation of such feedback systems it is necessary to first devise effective methods for capturing students' experience; this is the focus of the present paper.

Our work provides new insights on the role of physiological data collected from wrist-worn physiological sensing in monitoring learning experience. In particular, we make the following contributions:

- We present insights from a study that collects, during a typical class activity, physiological data from wrist-worn devices and standardized learning experience questionnaires.
- We show that physiological data have the capacity to monitor learning experience with acceptable accuracy.
- We show that with just a few minutes of data we can achieve a reliable estimate, and, thus, that real-time and ongoing monitoring is feasible.
- We discuss how our findings can democratize physiological sensing for self-monitoring, evaluation, and metacognitive reflection in learning.

2. Background and related work

2.1. Capturing students' learning experience

Experience is an episode, a chunk of time to remember; it is feelings and thoughts, motives and actions, all closely knitted together (Hassenzahl, 2010). Learning is an experience (Schmidt et al., 2019) that portrays students' feelings and thoughts, motives and actions when they interact with the teaching and learning environment (Biggs and Tang, 2007). Henrie et al. (2015) conducted a literature review on measurements related to learning experience and engagement, and summarized that quantitative self-report indexes are the most widely accepted and commonly used indexes to assess students' elements of learning engagement and experience. Such indexes are used by teaching and learning institutes to determine how to best use resources, people, and technology to engage students in meaningful and effective learning experiences.

To capture students' learning experience, several instruments have been used in the past; for example, the Course Experience Questionnaire (CEQ) (Ramsden, 1991) is a widely used instrument to assess learning experience after a given course or study program. CEQ assesses different aspects of student experience at the level of the whole course or program, and comprises scales related to quality of teaching, clarity of goals, and assessment. Another widely used instrument is the Virtual Course Flow Measure (Shin, 2006), which was conceptualized as a complex, multidimensional, reflective construct to capture flow experience during learning. Flow experience is the optimal state that people experience when engaged in an activity that is appropriately designed (Csíkszentmihályi, 2008).

In research on learning and instruction, various measures of mental or cognitive learning (e.g., Van Gog et al., 2012; Kay and Knaack, 2009) are regularly applied. Such measurements can reveal important additional information to researchers that is not necessarily reflected by more common performance measures, such as correctness of a task, speed, or number/type of errors made. Particularly, the combination of behavioral, cognitive, and emotional measures can provide information concerning the relative efficiency of training methods, in terms of the knowledge acquisition process and the quality of experience (see, e.g., Van Gog and Paas, 2008; Henrie et al., 2015; Kay and Knaack, 2009).

The multitude of assessment scales that have evolved to date do not provide a coherent, commonly accepted model for measuring learning experience (Van Gog et al., 2012; Henrie et al., 2015). A common practice that provides a more reliable and valid evaluation is to employ multiple assessment scales (Kay and Knaack, 2009; Van Zele et al. 2003). This approach leads to triangulation of data analysis and should be encouraged in both research and practice (Kay and Knaack, 2009; Skuballa et al., 2018). Thus, following suggestions from the literature (Henrie et al., 2015) that has categorized assessment scales into behavioral, cognitive, and emotional, we selected for our study three different constructs that have been found important in these three different categories: student usefulness and persistence in learning (Sánchez and Huero, 2010; Liaw and Huang, 2013; Fredricks et al., 2016), satisfaction with the learning activity (Gray and Diloreto, 2016; Liaw and Huang, 2013; Filak and Sheldon, 2008), and achievement/performance (Kuh et al., 2011; Kuvaas, 2006).

2.2. The quality of human-labeled data in learning and the potential of mobile sensing

Research that improves the design of teaching and learning needs measures of student learning experience that are reliable and accurate, but also easy to gather and scale, in order to evaluate the efficacy of the various learning design elements (e.g., instruction, learning materials). Reviews of measurement methods have identified issues that need to be addressed to improve the measurement of student learning (Fredricks and McColskey, 2012; Samuelsen, 2012). As mentioned above, such reviews have focused on self-report measures (questionnaires) of student learning, such as quantitative scales, after a given learning experience. However, questionnaires are not always the best method for measuring student learning experience. For instance, questionnaires can be inappropriate for primary school students (or other populations that are younger or have special needs), who may not fully understand the questions. Contemporary research and practice in learning (e.g., field studies, assessment) has considered self-reports as an objective measurement. However, using human-labeled data as ground truth entails certain limitations and threats to validity. For example, Van Gog et al. (2012) found that repeatedly measuring learning effort (using self-reported rating scales and associating them with response times) after performing an individual activity in a series favors longer activities.

Today, the collection of both human contributions and sensor data collected via mobile devices is becoming increasingly common across a range of contexts and methodologies. Despite their tremendous potential (e.g., collecting data in authentic settings) mobile self-report data

are not always accurate (Ickin et al., 2012; van Berkel et al., 2018b) and depends on various factors (e.g., context, participant fatigue) (van Berkel et al., 2019b, 2018a). Recent research in the area has identified methodological practices and techniques to improve the quality of human-labeled data (van Berkel et al., 2019a). In the area of learning experience, human-labeled data are considered as ground truth; however, their accuracy is an important area of research. Learning researchers tend to apply the scales either multiple times during the activity or once at the end (Van Gog et al., 2012). Asking students to provide a rating immediately after the activity requires them to reflect only on the activity they just finished, which is probably still (partly) activated in working memory, and returns more accurate responses (Raaijmakers et al., 2017). On the other hand, when asking students to rate only once, at the end of an entire course, and provide an overall rating requires learners to provide a retrospective judgment of the experience (e.g., the whole sequence of tasks), which has to be retrieved from long-term memory and might not be very accurate (Van Gog et al., 2012).

Thus, differences in the frequency with which learning rating scales are applied have a significant difference on the accuracy of the rating (Raaijmakers et al., 2017). Based on the literature (Van Gog et al., 2012), there are two theoretical reasons for favoring measuring learning experience immediately after each concrete learning task over conducting a single measurement at the end of a series of tasks. The first reason pertains to the usefulness of the measures. Learning experience is measured in order to compare different instructional approaches and designs, where such designs contain many different elements (e.g., lecturing and solving problems). When measuring elements related to learning only once at the end of the whole instructional process (or even course), it is impossible to determine the extent to which different elements contributed to the overall experience. Second, it is easier and for participants to recall and rate their experience after each activity, compared to measuring at the end of a series of instructional elements. It has been proven (Raaijmakers et al., 2017; Van Gog et al., 2012) that participants can accurately reflect on the instruction of an subtask they just finished, as it is still (partly) activated in their working memory. However, when learning is measured only once at the end of the whole instructional process, the rating will probably also involve information retrieved from long-term memory (Van Gog et al., 2012). Thus, it is unclear whether students estimate their learning based on an average of all instructional elements, the last elements they worked on, the most complex ones they worked on, or any combination of those possibilities.

Temporal data about learning are extremely useful but also difficult to obtain, especially with the limitations introduced by the use of surveys. Temporal data can give us momentary and continuous quantification of learning experience. Such temporal insights can inform the instructor so as to support the real-time management of learning activities (referred to as classroom orchestration [Dillenbourg, 2013]). As Henrie et al. (2015) stated, advanced mobile sensing technologies have the potential to contribute in new ways to capturing learning. In particular, in their recent literature review, Henrie et al. (2015) concluded that physiological sensors are effective for assessing students' parameters, but further work is needed to determine the type of information that needs to be collected and how this information will relate to contemporary standardized indexes. In this study, we utilize repeated, after-task, multi-item learning rating scales to capture learning experience, and then investigate the potential of parameters taken from mobile sensing technology to accurately predict the learning experience. Exploring the possibilities (and limitations) introduced by mobile sensing allows us to develop methods to increase data quality (e.g., accuracy, temporality) taken from human and sensor contributions.

2.3. Wearables and quantified-self technologies in learning

The confluence of physiological sensing and the quantified-self movement has the potential to provide authentic and granular learning activity data that will allow us to inform students and instructors (Lee et al., 2016). Previous works (e.g., Prieto et al., 2017; Papavlasopoulou et al., 2018; Sharma et al., 2019a) provided evidence of the value and usefulness of physiological data sources, including eye-tracking, facial-feature, and skin-conductance data (e.g., HR, blood volume pressure [BVP], electrodermal activity [EDA], and skin temperature). Various combinations of such data sources have been used in the past to explain (Raca and Dillenbourg, 2014) and/or predict (Beardsley et al., 2018) learning behaviors and/or performance (Junokas et al., 2018). Recent studies (e.g., Giannakos et al., 2019; Prieto et al., 2017; Sharma et al., 2019b) have provided strong evidence that physiological data sources (both wearable and stable devices) can provide an important source of information to explain different aspects of the learning experience.

Human-centered perspectives focus on the different states (e.g., cognitive, affective, and motivational) of the student at the moment of learning, and are best captured with fine-grained physiological and behavioral measures (e.g., electrodermal activity, facial expressions, actions) (D'Mello et al., 2017). Most of the literature has focused on physiological measurements coming from the autonomic nervous system, which can be measured more cheaply, quickly, and unobtrusively, and in a more ecologically valid manner, compared to those of the central nervous system. In the same vein, recent studies have found connections between physiological measurements and cognitive and affective states (e.g., mental workload, emotional valence) with the use of EDA (Ahonen et al., 2018). In the traditional classroom setting, Pijeira-Díaz et al. (2018) successfully utilized EDA, galvanic skin conductance, temperature, and accelerometer data to measure simultaneous arousal levels among students with respect to the students' mood, motivation, affect, and collaborative engagement.

During recent years, there has been increasing interest in wrist-worn sensing devices across a range of areas, including fitness, medical, entertainment, gaming, and lifestyle. The Vandrico Wearable Technologies database (http://vandrico.com/wearables) includes more than 400 wearable devices coming from more than 250 companies. It is evident that so-called quantified-self technologies are steadily gaining an important space in our life. These technologies assist people to collect personal data about their own behaviors, habits, and thoughts, with the aim of involving these users in self-monitoring and self-reflection processes to regulate different aspects of their own life (Ruiz et al., 2016).

Besides advantages such as the fact that quantified-self technologies are noninvasive, do not require much effort, and enable high-frequency user activity tracking, these devices provide a means to foster strong awareness, motivation, and behavioral change, with lifestyle and fitness trackers being common examples (Arnold et al., 2017). For instance, fitness trackers provide the wearer with information about how much activity they have undertaken, and allows them to set their own goal. Taking the analogy of a fitness tracker and applying it to learning settings has the potential to transform how students learn and instructors teach, as well as informing the various learning design decisions.

The literature has mentioned several advantages of quantified-self technologies in learning (Eynon, 2015). For example, quantified-self technologies can assist students to enhance their motivation, make informed learning choices, and enhance their metacognitive processes (Eynon, 2015). Another important potential implication of quantified-self technologies in learning pertains to identifying which typical school (and university) routines are associated with learning (Lee et al., 2016). For example, say that a student wants to see how their learning experience changes between hours of the day, periods (e.g., beginning of the semester, before exams), types of instruction, etc. Such insights cannot be obtained using traditional instruments applied to assess students' learning experience (e.g., surveys); conversely, quantified-self technologies offer this capability.

Despite promising findings from recent research studies (Di Lascio et al., 2017; Di Mitri et al., 2016; Pijeira-Díaz et al., 2016), there is

currently limited understanding of the ways in which quantified-self technologies can offer new insights into students' learning qualities (Wang and Cesar, 2015), which entails a risk of limiting creativity and learning approaches that might have long-term potential (Eynon, 2015; Selwyn, 2015). Utilization of sensor data in education is still a relatively new area, which needs more careful and critical attention from the respective research communities (e.g., Learning Analytics and Knowledge, Educational Data Mining, and Artificial Intelligence for Education). The increasing focus on quantified-self technologies in learning has raised numerous questions to explore that are both interesting and challenging. Understanding the potential and pitfalls of sensor data, and investigating them through the lens of widely used, standardized human-labeled data collection methods offers promising ways to assess learning qualities and inform the design of meaningful learning experiences.

The collection of student data via quantified-self apps can be transformative for students, especially those who are already familiar with activity-tracking mobile applications and quantified-self technologies. Thus, we propose that research should pursue multi-pronged approaches and the collection of activity data, as well as investigating the extent to which those data can provide insights into students' learning experience in an accurate and timely manner.

3. Method

3.1. Context

To capture fine-grained physiological data during a class activity, we conducted an in-the-wild study. The study took place in the context of a university course called *Customer Driven Project*. This is a master's-level class in which groups of 5–6 students (in the first year of their master's degree) work on a software engineering project with a real customer. For each group, a different stakeholder has the role of the customer and allocates a software product to be developed throughout the semester. The final goal for the groups is to deliver a functional prototype/product and a project report, together with a short video that describes the developed product and the group's work. During the semester, groups have internal group meetings, meetings with the customer representative, and class sessions with their advisor (i.e., the instructor of the group). The role of the advisor is to meet with the groups once each week for approximately 30 min in order to discuss with them the progress and objectives of their project. Specifically, the advisor follows the project's progression, ensures that sufficient contact with the customer is maintained, coaches the team, helps with challenges that arise during the overall process, and gives feedback on the group's project report. During the weekly class sessions, based on the group's needs, discussions are conducted with all group members (see snapshot from a class session in Fig. 1, right). At the end of each class session, students are asked to rate their experience based on a standardized questionnaire that is designed to capture their learning during the session (Fig. 1, left).

3.2. Participants

We recruited 31 university students (12 female and 19 male) aged between 21 and 53 years old (mean = 24.01, S.D. = 5.87) forming 6 groups (five groups with 5 members and one group with 6). Participants were recruited using convenience sampling from the pool of a major European university. Participants were CS majors and taking the respective course as part of their CS degree. Participants were given a 30 euro gift card upon completion of the study.

3.3. Procedure and apparatus

At the beginning of each class session, once the participants had sat down at their respective places, they put on an Empatica E4 wristband and attended the class as usual. At the end of the class, each participant completed a questionnaire that was designed (based on the literature) to capture their learning experience. The class lasted approximately 35 min (mean = 34.26 min, S.D. = 6.4 min), with approximately 25 min being the actual class activity (mean = 25.1 min, S.D. = 13.1).

Empatica's E4 wristband is considered one of the most popular and validated systems, along with UFI's model 1020 Pulse plethysmograph, and Biopac's Bionomadix PPGED-R add-on to its MP150 system. The E4 wristband is based on the E3 band, which has been measured up to the ECG gold standard (Greene et al., 2016). Studies have conducted comparisons between the different devices, showing similar quality of data (Empatica, Inc.; Barreto et al., 2007; Garbarino et al., 2014). Our decision to use E4 wristbands was motivated by the fact that E4 captures all data streams at the same time with high accuracy (Greene et al., 2016) (i.e., device error = 0.2 ˚C within 36–39 ˚C for temperature; 1% on EDA and least count for Photoplethysmography sensor [BVP] is 0.9 nano watts) and the least intrusion and requirement of maintenance (Greene et al., 2016).

3.4. Experimental design

The research design of our study comprised a single-group time-series design (Ross and Morrison, 2004) with both continuous (physiological data) and post (questionnaire) measurements of each group, with the experimental treatment induced. Each participant was recorded three times (one class per week for three weeks), and recorded in three different class activities lasting, on average, 25.1 min (S.D. = 13.1). Fig. 2 presents the protocol of our experiment. Each participant spent 3–4 min sitting down, putting on the wristband, and initiating it. The class activity then took place and the participant finally spent 5–6 min completing the questionnaire (Fig. 2). Each session (set-up, class activity, and questionnaire time) lasted 34.26 min (S.D. = 6.4).

3.5. Measures

At the end of each class session, students completed a paper-based survey. The surveys gathered feedback regarding students' learning experience. The survey concerned factors adopted from prior studies, particularly on how students perceived the following three notions: (1) Satisfaction (SAT), (2) Usefulness (USE), and (3) Performance (PER). Table 1 summarizes the operational definitions of these factors, the items/questions, and their respective bibliographic sources. The selection of these three attitudinal factors and the respective measures were based on literature and prior studies. To this end, the three factors (and the respective 10 questions) were used to quantify students' learning experience at the end of each class activity. In all measures, a 7-point Likert scale was applied (1 = "not at all" – 7 = "very much").

Physiological parameters have been closely associated with emotional and cognitive processing (e.g., Cowley et al., 2013; Di Lascio et al., 2017; Di Mitri et al., 2016). Implicit emotional responses that may occur unconsciously, such as threat, anticipation, salience, and novelty, can be examined using EDA. In addition, certain emotions trigger the release of hormones such as epinephrine, which increases blood flow to bring more oxygen to the muscles (and increases BVP). The change in blood volume is proportional to the HR and peripheral temperature. Such responses have been used to infer various learning-related constructs, including cognitive load (Hussain et al., 2013), perceived difficulty (Pham and Wang, 2016), and learning performance (Cowley et al., 2013), and are also used in this study.

In particular, during the class sessions we captured data on participants' physiological parameters using the Empatica E4 wristbands, which include sensors for HR, blood pressure, temperature, and EDA levels. Participants wore the wristband on their nondominant hand, and four different measurements were recorded: (1) HR at 1 Hz, (2) EDA at 4 Hz, (3) temperature (TEMP) at 4 Hz, and (4) BVP at 64 Hz.

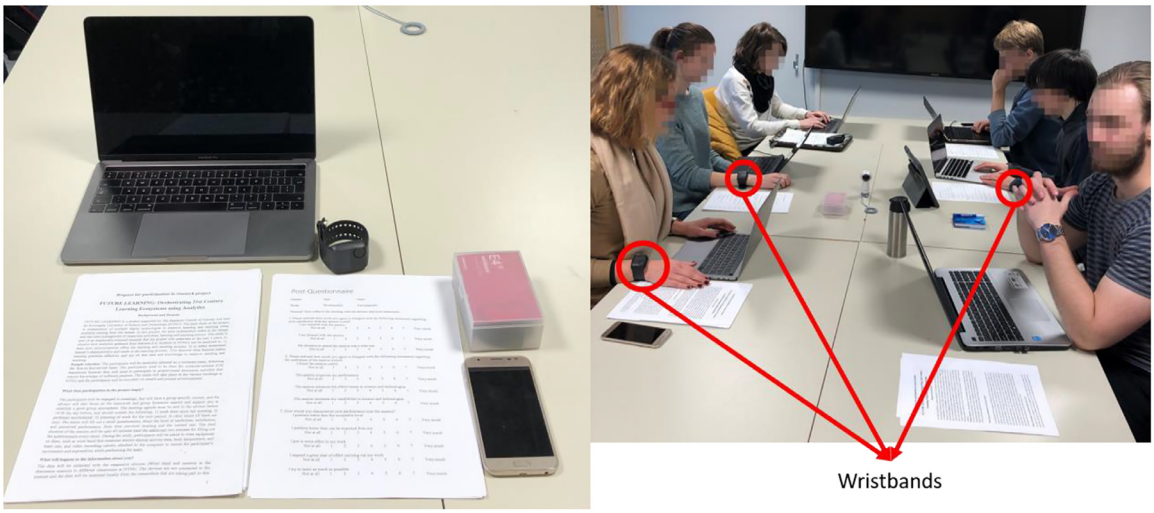

Fig. 1. Data collection setup (left) and a snapshot of the class session (right).

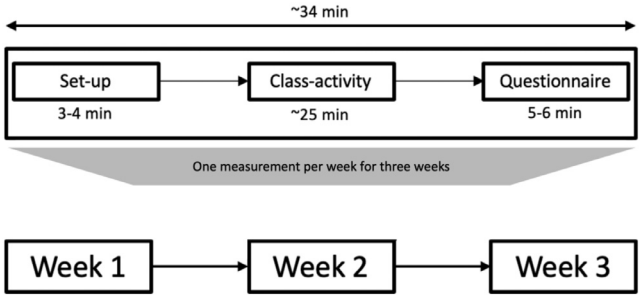

Fig. 2. Protocol of the study.

4. Results

4.1. Wrist data preprocessing and feature extraction

Empatica data can be affected by age, gender, time of the day, and other physiological conditions. To remove personal and other conditional biases from the time series data, we normalized the time series as a proportion of the mean of the first 10 s of the data. Further, we used a MinMax normalization to have all the data range from 0 to 1. Further, to remove noise from the Empatica E4 data sources (EDA, HR, BVP, TEMP), we fit a spline curve on the time series data. Fig. 3 shows the steps taken to proceed from the original time series to the smoothed one.

From the smooth time series, we followed related work in utilizing sensor data to understand learning phenomena (e.g., Prieto et al., 2018; Di Lascio et al., 2018; Giannakos et al., 2019). In particular, we computed the first five auto-correlation coefficients, as proposed by Box et al. (2015), and further utilized in classification of mental tasks by later work (Rahman et al., 2018). 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 computed the Fourier transform of the electrode signals and took the first five coefficients (first five dominant frequencies) (Sitnikova et al., 2009). The energy of a sensor's signal has been proposed to be a valid and reliable indicator of mental fatigue (Xu et al., 2018). Moreover, histogram-based features, such as max, skewness, and kurtosis, are also important predictors of learning experience and engagement (Prieto et al., 2018; Di Lascio et al., 2018; Giannakos et al., 2019). Thus, building on previous related works, we extracted the following features:

1 Histogram-based features—min., max., mean, median, S.D., skewness, kurtosis
2 Fourier transform—first five coefficients
3 Energy of the signal—root mean square of the signal
4 Autocorrelation—first five coefficients

This gave us a total of 18 features per data stream. Before

Table 1
The factors and their respective items/questions used in our study.

Factor (source adapted from)	Operational definition	Questions/items
Satisfaction (Gray and Diloreto, 2016; Liaw and Huang, 2013)	The degree to which a person feels positively about the activity.	I am satisfied with the session (SAT1) I am pleased with the session (SAT2) My decision to attend the session was a wise one (SAT2)
Usefulness (Sánchez and Huero, 2010; Liaw and Huang, 2013)	The degree to which an individual believes that attending the respective session is useful for him/her.	The session improved my performance (USE1) The session enhanced the effectiveness in science and technologies (USE2) The session increased my capabilities in science and technologies (USE3)
Performance (Kuvaas, 2006)	Students' rating of their performance.	I performed better than the acceptable level (PER1) I performed better than is typically expected from me (PER2) I put extra effort into my work (PER3) I expended a great deal of effort carrying out my work (PER4)

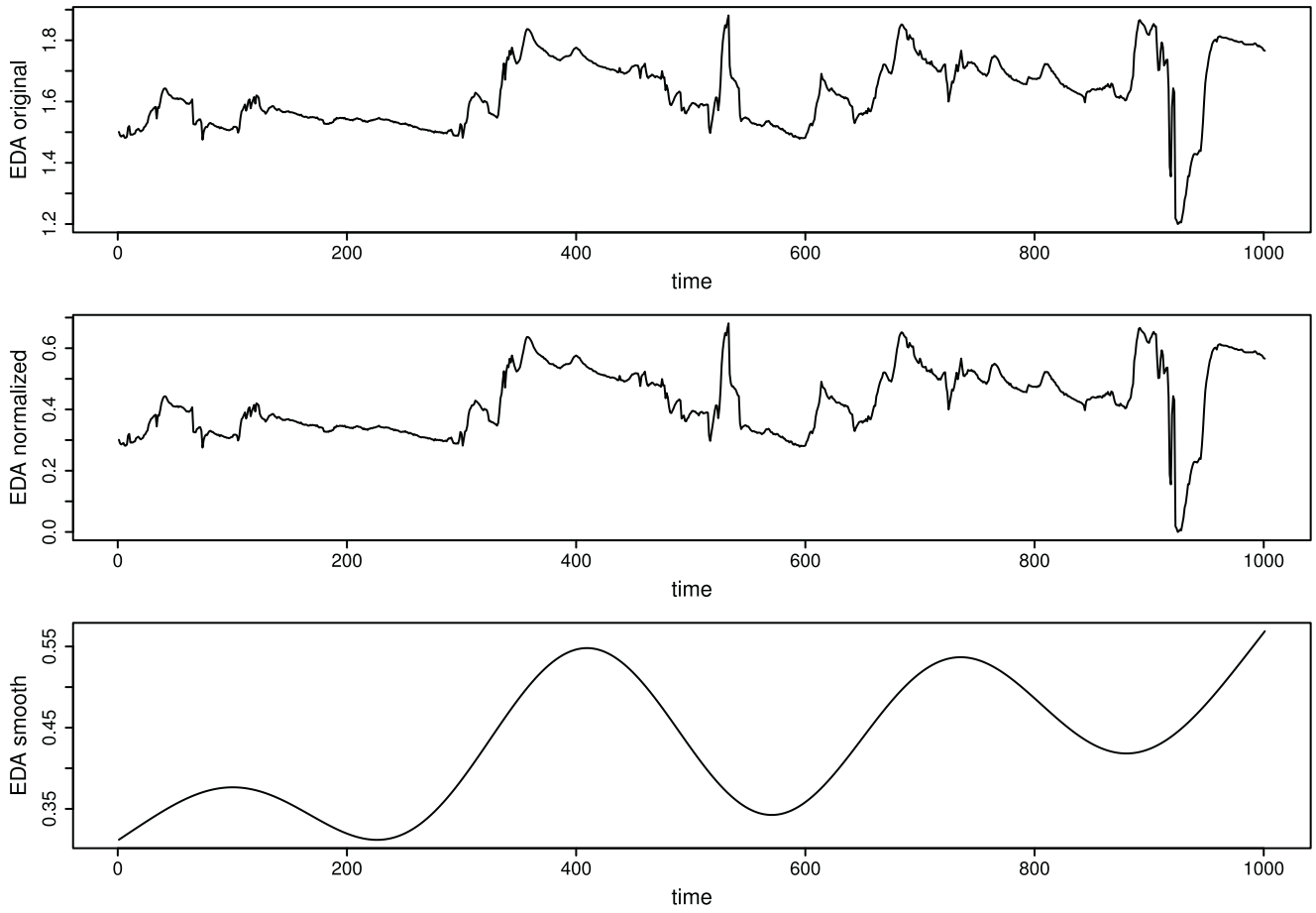

Fig. 3. Exemplar representation of the data preprocessing steps.

proceeding with the noise-removal and the feature-extraction processes, the data corresponding to the first two minutes from each individual were removed to ensure that initial spikes due to any kind of unwanted additional movements were removed.

Noise, or "artifacts," can be introduced whenever an individual adjusts the sensor, knocks the wearable against something, or places pressure on the device. We used EDA Explorer (Taylor et al., 2015)—which is a machine learning classifier, focusing mainly on using support vector machines, that detects noise with 95% accuracy—to remove artifacts.

4.2. Reliability and validity of the questionnaire measures

Fornell and Larcker (1981) proposed three procedures to assess the convergent validity of any measure in a study: (1) the composite reliability of each construct, (2) the item reliability of the measure, and (3) the average variance extracted (AVE).

Thus, we first carried out an analysis of composite reliability and dimensionality to check the validity of the scales used in the questionnaire. Regarding the reliability of the scales, Cronbach's alpha (α) indicators were applied (Cronbach, 1951). As Table 2 shows, the result of the test revealed acceptable indices of internal consistency in all factors.

In the next stage, we proceeded to evaluate the reliability of the measures. The reliability of each item was assessed by measuring its factor loading onto the underlying construct. Hair et al. (2006) recommended a factor loading of 0.7 to be good indicator of validity at the item level. The factor analysis identified three distinct factors: (1) satisfaction, (2) usefulness, and (3) performance (Table 2).

The third step for assessing the convergent validity is to assess the AVE. The AVE measures the overall amount of variance that is attributed to the construct in relation to the amount of variance attributable to measurement error. Convergent validity was found to be adequate when the AVE is equal or exceeds 0.50 (Segars, 1997).

Table 2
Summary of questionnaires' measurement scales.

Factors	Questions/items	Mean	S.D.	Loadings	α	AVE
Satisfaction	SAT1	6.06	0.93	0.881	0.830	0.68
	SAT2	5.93	1.05	0.864		
	SAT3	6.11	1.09	0.716		
Usefulness	USE1	4.74	1.42	0.769	0.833	0.68
	USE2	4.28	1.21	0.798		
	USE3	4.04	1.23	0.909		
Performance	PER1	4.54	0.95	0.793	0.804	0.62
	PER2	4.23	0.89	0.811		
	PER3	4.56	1.26	0.799		
	PER4	4.90	1.27	0.747		

Notes: SD = standard deviation; α = Cronbach's α; AVE = average variance extracted.

In order to identify the correlations among the three attitudinal variables, we used Pearson's correlation coefficient, which quantifies the strength of the relationship between variables. Pearson's test verified the relatively strong relation among the three factors, as indicated in Table 3. In addition, the variables were tested for discriminant validity, which requires the square root of the AVE of each variable to be larger than its correlation with the rest of the variables (Table 3).

Table 3
Pearson's correlation coefficient between factors.

	Satisfaction	Usefulness	Performance	Square root of AVE
Satisfaction	1			0.82
Usefulness	**0.530****	1		0.82
Performance	**0.289****	**0.250***	1	0.78

Notes: All correlations are significant.
** $p < 0.01$.
* $p < 0.05$.

4.3. Feasibility of the learning experience prediction

We considered all possible combinations of the four data streams (16 in total). For each combination, we used four different training algorithms (random forest [RF], SVM with linear, radial and polynomial kernels) to calculate RMSE in relation to the three questionnaire factors (SAT, USE, and PER). The selection of the four training algorithms was in accordance with the related work in ubiquitous computing (see, e.g., Di Lascio et al., 2018), but also due to the qualities of the algorithms. In particular, all four training algorithms can be used to predict both the categorical and the continuous target variables. Hence, using them is an acceptable way to obtain a certain level of generalizability in case other researchers want to use the features and algorithms with other categorical target values. SVMs allow the addition of latent dimensions to provide better separability in the feature space. This is necessary to obtain a lower amount of error, especially when there are not many individuals and the number of features approaches the number of individuals (in our case, 31 individuals and 18 features per data stream). SVMs are also useful in cases where there is unbalanced data, as in our case where the target values are skewed towards the higher values. The different kernels (linear, polynomial, and radial) are mostly used in an empirical manner to introduce the latent dimensions. RF is an addition that provides researchers with a way to tackle missing data. In cases where the data are noisier than a certain level (this does not apply in our case), RF is known to maintain the accuracy of a large proportion of data. In addition, RF prevents the overfitting of training data by not allowing more than a certain number of trees in the model. Such an algorithm is useful in cases where there are few individual samples. Thus, we elected to utilize the aforementioned training algorithms due to their advantages, but also to increase the generalizability, reusability, and potential future comparison of our results.

We employed leave-one-subject-out testing, with the outcome indicating that the models were not overfitted using the training data (see Appendix B). Fig. 4 shows the comparison of the different training algorithms and different data streams (BVP, HR, EDA, TEMP), with SVM with polynomial kernel found to outperform other algorithms. The results show that the minimal error rate achieved is 13.8% for USE, 11.0% for PER, and 11.8% for SAT (Fig. 4). All errors, on a 7-point Likert scale, translate to less than 1 point.

To quantify how well our models outperform a random model, we performed a random guess prediction on the three dependent variables obtained from the survey. The random guess models (in Table 4) confirm that the error rates of the optimal predictions are significantly below the random-baseline.

Overall, we find that SVM with polynomial kernel outperforms the other training algorithms. The pairwise differences between SVM polynomial and SVM radial indicates that there is no statistically significant difference (t-tests show p-values greater than 0.05). RF significantly underperforms when compared to SVM radial and SVM polynomial (t-tests show p-values less than 0.05). Finally, SVM linear was found to statistically underperform the other three training algorithms (t-tests show p-values less than 0.001). The detailed results can be found in Table 5.

Next, we consider in detail the algorithm that gives the best results (i.e., SVM with polynomial kernel) and the combinations of the data streams to calculate RMSE for the three dependent variables (Fig. 5). We observe that there is no significant prediction power of any of the four data streams (i.e., HR, EDA, TEMP, BVP) independently or in any possible combination. Looking at the features calculated from those data streams (Appendix A), we observe that the best predictors for the three dependent variables are different from each other. First, for satisfaction, the best predictors are: the most dominant frequency of HR, the variance in blood pressure, and mean temperature and EDA. Second, for performance, we observe that the temporal features (autocorrelation coefficients) from HR and BVP are among the most valuable predictors. Finally, temperature-based features appear to be the most important among the top 10 predictors in predicting the perceived usefulness of the sessions.

4.4. Time needed for predicting the learning experience

To identify the minimum time needed to achieve a reliable estimation of learning experience, we attempted to predict learning experience using different segments of the data (e.g., 100%, 50%, 25%, and 12.5%) until we observed a significant increase in the error rate. For example, “using 25% of data” translates into dividing the data into four quarters based on their length of time and using only the first quarter for prediction purposes.

SVM (polynomial and radial) had the most accurate prediction overall. Running the analysis (with both the top-performing algorithms) with 50% and 25% of the data, we did not observe any significant difference compared to 100% of data, while when we used 12.5% of our data there was a significant increase in RMSE (see Table 6). Nevertheless, even with 12.5% (approx. three minutes) of data, SVM polynomial outperformed the random baseline. For each of the segments we used a rolling window with an overlap of 50%. For example, considering the quarters of the data, we first applied the prediction to the first quarter of the data, then shifted the window by one eighth of the length and took the further quarter length. This resulted in seven segments of data with a length equal to 25% of the total length. Similarly, we obtained three halves and 15 eighths of the data.

Thus, when we moved from 12.5% to 25% length of data, we noticed a significant decrement in RMSE, with 25% of the data providing 14.1–14.3 average RMSE; adding more data did not improve the performance of the classifiers. This could be because of the relatively homogenous level of students’ experience and/or the relatively short duration of the activity (mean duration 25.1 min, S.D. = 13.1 min). Thus, the “mood” was set early in the class activity and there were no significant changes. Therefore, using time equivalent to 25% (6.1 min, S.D. = 3.3 min) of the data provided an accurate prediction in our sample (i.e., low RMSE values: mean = 13.8%, S.D. = 3.5%). On a 7-point Likert-scale, this error translates to less than a one-point absolute difference.

5. Discussion

Our results suggest that physiological data obtained from wearable sensors can be a proxy for the learning experience. In particular, our findings indicate that physiological data coupled with machine learning algorithms give us a relatively good estimation of the learning experience (i.e., 11% error). Our results confirm those of Henrie et al. (2015) who reviewed measurements related to learning experience and engagement and indicated the potential of sensing technologies to contribute in new ways to capturing learning. In a recent literature review, Mangaroska and Giannakos (2018) indicated that only a few studies so far have utilized sensing technologies to inform learning design. Our results highlight the potential of noninvasive sensing technologies to inform learning design, and even complement real-time insights drawn from alternative student-generated data, such as log traces (Pham and

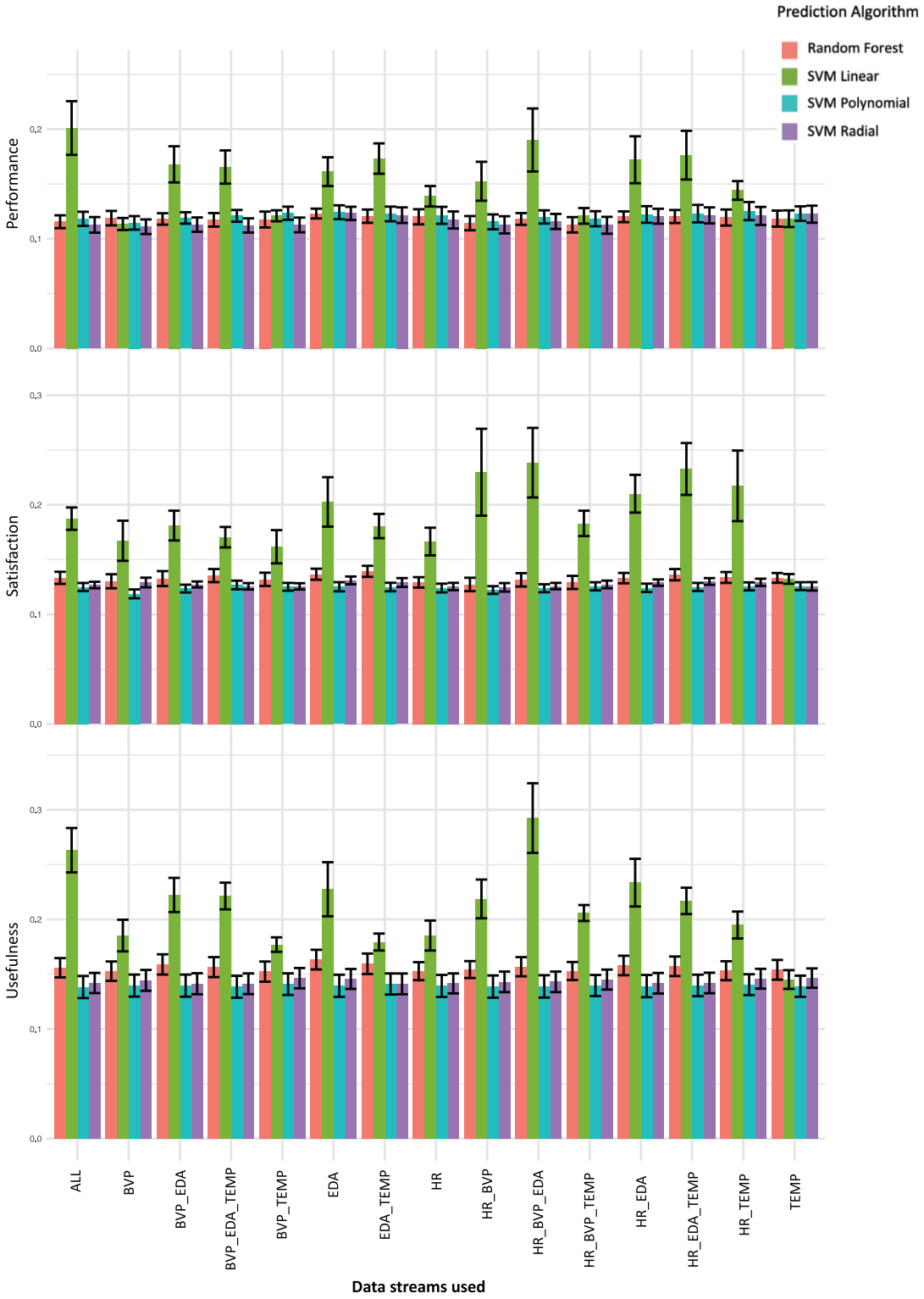

Fig. 4. All combinations of the data streams and the different algorithms to calculate RMSEs (shown on y-axis).

Table 4
Random guess baseline RMSE of the three dependent variables used in the survey.

Dependent variable	Performance	Satisfaction	Usefulness
Random guess baseline RMSE	0.279	0.396	0.292

Wang, 2016).

Considering the evidence in the literature regarding which measurements relate to learning experience and engagement (Henrie et al., 2015), it is apparent that quantitative self-report indexes are dominant today. This is the case despite the fact that there are various disadvantages of self-report indexes (e.g., continuous measurements, temporal insights). Furthermore, the effectiveness of contemporary experience measures that can be obtained implicitly using devices such

Table 5
Results of pairwise difference t-tests between SVM-polynomial, SVM-radial, SVM-linear, and RF.

Pair	T-value	Degrees of freedom	p-value
SVM polynomial and SVM radial	−1.32	60	.62
SVM polynomial and SVM linear	−9.89	60	<0.001
SVM polynomial and RF	−3.34	60	<0.05
SVM radial and SVM linear	−9.36	60	<0.001
SVM radial and RF	−2.24	60	<0.05
SVM linear and RF	8.19	60	<0.001

as electroencephalography (EEG) and low-cost Web cameras has been proven, and these devices have even provided additional promising affordances (Sharma et al., 2019a; Hassib et al., 2017; Whitehill et al., 2014; Monkaresi et al., 2017). Taking into consideration the practical and technical difficulties, as well as the high cost, of utilizing sensing devices such as EEG, it is clear that wearables and quantified-self technologies can provide a good solution for collecting insights about learning (Lee et al., 2016; Arnold et al., 2017). Our results provide evidence that wearable sensing technologies offer accurate measurements about the learning experience and have the capacity to support quantified-self technologies to strengthen students' self-regulated learning and self-responsibility. We also highlight the importance of obtaining knowledge about the self, and offer several socioeconomic implications regarding the way in which the individual student and the teacher can use those insights and improve themselves (Lupton, 2014).

Our results (see Fig. 5) indicate that we can infer students' learning experience with good accuracy from any of the four data streams (i.e., HR, EDA, TEMP, BVP). Thus, researchers do not need to invest in expensive equipment and procedures to scale up and democratize sensing-based student experience. Any of the market devices, such as the Fitbit, Jawbone's Up, etc., provide a reasonable subset of the needed data streams. Moreover, although SVM polynomial was found to be the most accurate algorithm, SVM radial and RF algorithms also provided accurate predictions. The only exception is SVM with linear kernels, which (although providing results significantly higher than the baseline) produced significantly less accurate predictions. This indicates that we need more sophisticated classification prediction than a linear classification, but, at the same time, there is no need to dig into very complex classifiers.

Our results regarding the time needed to provide an accurate estimation of learning experience show that with a few minutes of data (6.1 min, S.D. = 3.3 min) we can obtain a relatively accurate prediction (i.e., low RMSE values: Mean = 13.8%, S.D. = 0.1%). This might be subject to the type of learning activity and the homogeneity of students' learning experience; however, it provides evidence that we can make early estimations based on students' physiological data. This opens new avenues of research on evaluating different short learning experiences, and performing A/B experiments in relatively short learning activities—even within the same lecture hour or other class activity.

In accordance with the results of previous works (e.g., Sharma et al., 2019a; Pijeira-Díaz et al., 2018), the results of our study support the value of physiological data sources, and in particular demonstrate how skin conductance data can offer continuous and unobtrusive monitoring of learning. Previous studies have demonstrated that gaze data is a key element in understanding and predicting learner behavior and/or performance due to the direct connection with students' mental effort (Bednarik, 2012; Kaller et al., 2009). However, gaze data provide different information than skin conductance (e.g., users' attention and focus); thus, fusing multimodal physiological data sources has the potential to provide even more accurate predictions (Giannakos et al., 2019). Therefore, whenever possible, researchers need to leverage multimodal physiological data, and weigh up the trade-off in capacities (e.g., gaze, brain, face, skin) and limitations (e.g., ecology, cost) of the various modalities.

5.1. Theoretical implications for contemporary learning and instruction

Contemporary learning scenarios integrate individual activities (e.g., reading), team work (e.g., problem solving) and class-wide activities (e.g., lectures). Some of these activities are enhanced with technology, while others are not. In addition, some are face-to-face while others are online. For such learning scenarios to be successful, proper classroom orchestration is essential (Dillenbourg and Jermann, 2010). Classroom orchestration refers to the design and real-time management of multiple learning activities (Dillenbourg, 2013). The notion of classroom orchestration was coined to understand instructors' difficulties in adopting innovative technologies and practices in their teaching (Dillenbourg, 2013). The most important barrier in the adoption of innovative teaching practices and technologies is the orchestration load of the instructor (Dillenbourg et al., 2016). Orchestration load relates in part to the real-time identification of students' learning and guidance—a task that the instructor has to undertake for several group of students, or even for every student. The results of this study show that with just a few minutes of data we can achieve a reliable estimate of students' experience; thus, real-time and ongoing monitoring is feasible. Therefore, it is possible to track student learning experience across the activity; this result opens new avenues for developing technologies that collect and even visualize students' learning experience, to reinforce self-directed learning and reduce the orchestration load (Dillenbourg et al., 2016). In addition, this has various other implications, such as allowing teachers to reflect on their instruction and test/evaluate innovative practices and technologies in their classroom.

According to flow theory (Csikszentmihalyi, 1997), to support meaningful learning one should be in a flow state that neither frustrates them nor deters them from the activity. Flow is defined as the optimal state that people experience when engaged in an activity that is appropriately challenging to one's skill level, often resulting in immersion and concentrated focus on the task (Csíkszentmihályi, 2008). A student can experience relaxation in a learning activity when their skill level is very high and the activity challenge is very low. Conversely, a student can experience anxiety when their skill level is very low and the activity challenge very high (Csíkszentmihályi, 2008). Neither of the two states is supportive for learning. Being able to keep students in the flow learning experience (that is engaging but does not overload them) is essential for teachers to nurture. Thus, an indication of students' experience can support teachers' or students' decision regarding remedial action regarding their learning/instruction. The various instruments that exist—for instance, self-reports (Henrie et al., 2015) and the NASA task load index (Hart, 2006)—cannot account for the rapid changes of a learning activity, such as reading, following an instruction, solving a problem, working in teams, discussing, etc. Therefore, the utilization of wearable devices to capture students' engagement offers a solution that provides the necessary affordances (e.g., self-awareness and reflection) to support flow state in learning.

Smart learning is a new term that has come to describe technological and social developments that enable effective, efficient, engaging, and personalized learning (Giannakos et al., 2016). Collecting and combining learning analytics has the potential to provide valuable information for designing and developing smart learning environments. The concept of smart learning environments is enabled by technologies that rely on sensors and new ways of connecting and exchanging information (Spector, 2014); while the concept is of growing significance, scholarly work is lacking, both conceptually and empirically (Giannakos et al., 2016). Our study demonstrates how wearable technologies can provide important information about students' learning experience. Such Information can be utilized to support experience-aware technologies, since smart learning embraces the concepts of learning ecosystem and distributed information and intelligence (Hwang, 2014).

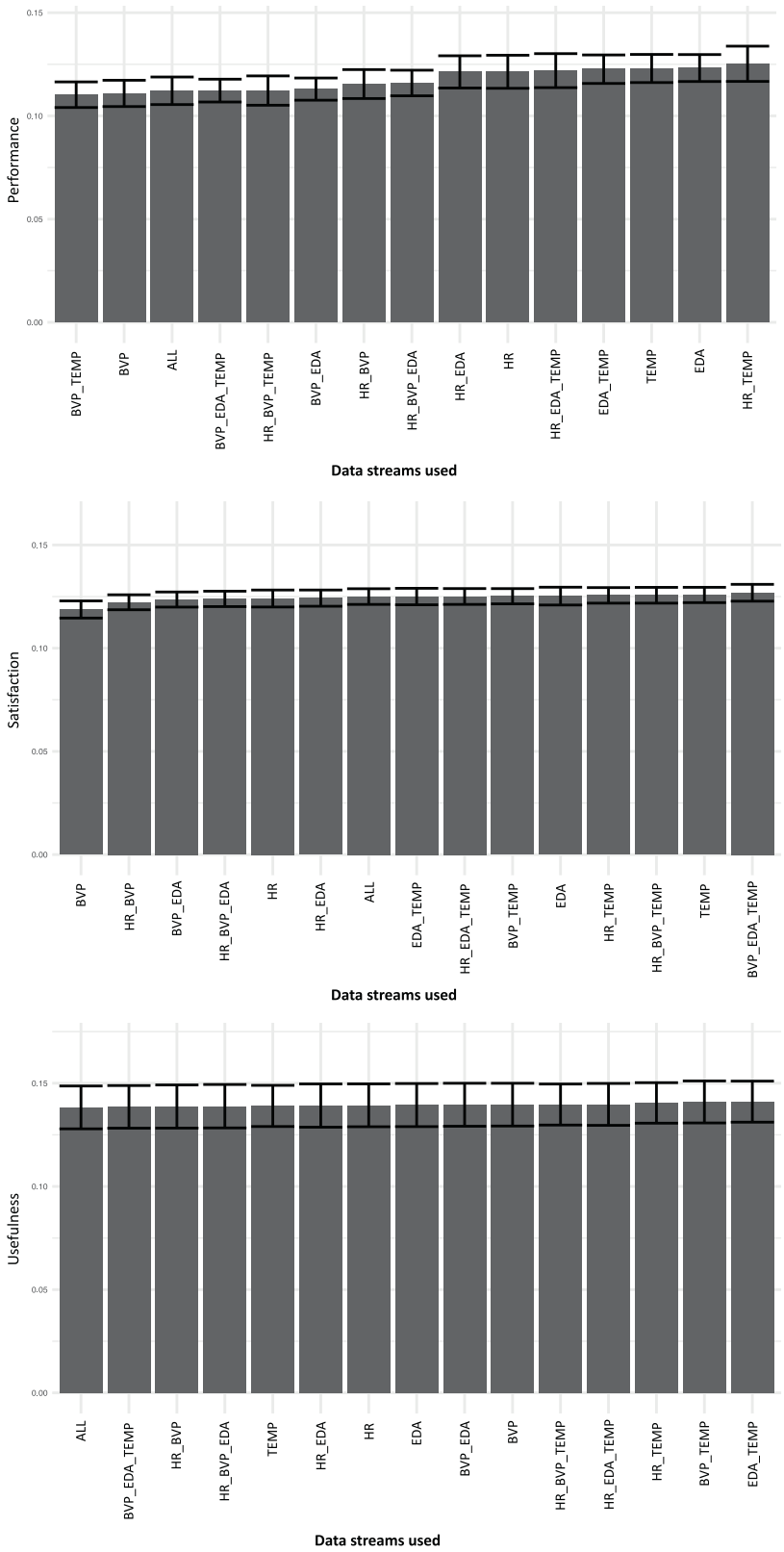

Fig. 5. Using the best predictor (i.e., SVM polynomial) with all combinations of the data streams to calculate RMSEs (shown on y-axis).

5.2. Practical implications

The possibility of using wearable sensing devices to automatically capture students' learning experience opens up new avenues for the quantified-self movement to support learning. First, it provides early feedback to both students and teachers. Thus, it makes it possible for teachers to get early warnings of poor mental effort, disengagement, and disinterest, and allows students and teachers to take remedial action very early. The physiological data we utilized have been found to be strong predictors of mental effort (Rahman et al., 2018), learning

Table 6
RMSE values using the different lengths of data segments, applying SVM polynomial and all data streams together.

Factors	Average RMSE values (%) with n moving windows			
	Whole (n = 1)	Halves (n = 3)	Quarters (n = 7)	Eighths (n = =15)
Performance	11.0 (1.9)	14.0 (4.1)	14.1 (3.2)	29.5 (6.5)
Satisfaction	11.8 (2.1)	13.8 (3.5)	14.3 (4.6)	27.8 (8.7)
Usefulness	13.8 (2.8)	13.9 (2.9)	14.1 (2.6)	25.7 (9.2)

Note: Numbers in column headers show the number of windows (with 50% overlap) used in the predictions.

engagement (Di Lascio et al., 2018), and learning outcome (Wang and Cesar, 2015); our study validates their capacity to monitor learning experience with acceptable accuracy, and shows that with just a few minutes of data we can achieve a reliable estimate. Thus, the practical implications of wearable sensing technologies in learning are greatly strengthened.

Our results have several implications for learning and HCI research. First, we provide evidence regarding the feasibility of utilizing sensing data collection and capturing students' learning experience. This allows researchers (but also enables teachers who want to experiment) to explore the effectiveness of different instruction and learning design decisions in a relatively easy and agile manner. Second, it provides a way to support student groups that are unable to provide questionnaire feedback (due to, for instance, a lack of language competence or attention deficits). Third, it allows us to identify potential ways in which to alter the learning experience during mini instructional phases during a class activity (e.g., solving exercises, discussing project work etc.), which cannot be investigated via traditional questionnaires when measuring learning only once, at the end of the whole instructional process (or even course).

Overall, the combination of wearable sensing technologies and quantified-self applications to support self-monitoring, evaluation, and metacognitive reflection in learning provides an important practical, but also theoretical, implication. In doing so, it opens promising avenues for future research—for instance, in practical aspects of learning design and pedagogy, as well as the integration of self-monitoring in learning models and theories. It also reveals how learning analytics can be applied in a more ubiquitous and agile way, going beyond classroom and formal learning settings and supporting responsive learning and self-improvement.

To inform the design of wearable devices to support learning, it is important to identify the most important data features for learning and develop algorithms and technologies that utilize them. Looking at the top 10 features (Appendix A) to predict the three dependent variables, we notice that histogram-based features (e.g., mean, kurtosis) are the most important predictors of the learning experience. Especially with respect to satisfaction and usefulness, seven out of the top 10 and six out of the top 10 features are histogram-based, respectively. For performance, we see that autocorrelation provides seven out of the top 10 features. Auto-correlation coefficients describe the correlation between the values of the same signal at different times as a function of the time lags (time domain), and has been found to be an important predictor in mental effort and mental tasks (Rahman et al., 2018). Thus, listing histogram-based features as the most important for emotional and behavioral aspects of the learning experience, and auto-correlation features as the most important for performance, validates previous works (Prieto et al., 2018; Di Lascio et al., 2018; Giannakos et al., 2019) and extends them by quantifying the importance of these aspects. Looking at the data streams of the top 10 features (see Appendix A), BVP, HR, and TEMP are the top ones. This is probably connected to the fact that certain emotions and types of engagement (e.g., interest, attention), increase blood flow and bring more oxygen to the muscles (and thereby increase BVP, which is proportional to HR and temperature). This is very interesting for the future design of wearable devices to support learning, since monitoring these physiological parameters is relatively easy with today's technologies and calculating these features can be done "on the fly," which allows us to inform students and teachers (or even contemporary learning systems) about momentary (temporal) learning experience. This opens up new opportunities for learning systems affordances (feedback mechanisms), as well as for advancing contemporary algorithms (e.g., adaptive algorithms, learner modeling, recommended systems for learning).

The design and development of such systems does not entail simply embedding a display in a wall or a table and providing visual analytics that can help the students or the teacher. For such a system to be used successfully, we need to invent communication channels that do not require full attention, but rather allow teachers to perceive information and perform physical actions in the background or periphery of attention (e.g., peripheral perception, peripheral interaction). Hence, further work is needed in classroom settings in order to identify ecologically valid ways to introduce those technological affordances to the teacher (e.g., information on a central display or transmitted via wearable devices), without hindering their cognitive abilities (e.g., split attention, high mental effort). In many cases, low-resolution information (e.g., average information about students' engagement and learning, rather than a list of their names and the associated indexes) provides students (and teachers) with some awareness of the state of some groups, some activities, some students etc. The notion of awareness reflects the influence of another close community, computer-supported cooperative work. Initially, awareness tools for learning were displayed on the computer screen, but since computer displays are cluttered with information, awareness tools started exploiting ubiquitous and ambient systems (e.g., peripheral displays, background sounds, vibrations of furniture, etc.). Therefore, further work is needed in order to investigate how awareness and reflection tools can incorporate insights collected from wearables and nonintrusive devices (e.g., cameras) during learning, and align those tools with teachers' practice and beliefs.

5.3. Limitations

The findings support our proposition that physiological data coming from wristbands have the capacity to provide insights into students' experience during a learning activity; however, the findings are also subject to certain limitations. The participants of our study were graduate students, representing an appropriate sample for our study since we wanted a student population that could effectively read the standardized survey and provide accurate responses. However, younger or older populations (e.g., those in primary education, lifelong learning, professional training, etc.) might produce slightly different results. To test our proposition, we conducted an in-the-wild study; such studies produce data of high ecological validity, but are vulnerable to potential disruptions and noise. In our case, however (i.e., a project-based learning course), such disruptions are very rare. In addition, participants were aware of the data collection since they had signed a detailed consent form, which may have led to increased desire "to provide good data." Nonetheless, controlling physiological data is not something students can easily do, especially for the 25-minute period of the class, and all three class activities. Moreover, the leave-one-subject-out cross-validation removed any such bias.

Our study was conducted in a project-based learning class lasting approximately 25 min, and these conditions also induced specific characteristics in our study. For instance, a continuous discussion was conducted among the students/instructor and it was difficult for any of these parties to become completely disengaged. In order to obtain more accurate information about students' experience the duration of the activity was relatively short, and this might also have impacted the results (e.g., a longer learning activity might have resulted in more variations in students' experience). Therefore, the generalizability of our findings is constrained by the learning activity and the context,

since longer passive lectures, hands-on lab exercises, or different dynamics and learning design might give different results. However, this study employed a learning activity that is widely used in the contemporary educational system (e.g., project-based learning, peer tutoring) and, due to its limited variations in learning experience and relatively short duration, can serve as a baseline for future studies.

Measuring learning experience involves inference, and inference involving complex psycho-physiological constructs involves a degree of error. This is irrespective of whether the inference is performed by a human or a machine, and in some cases the computer even outperforms humans (D'Mello et al., 2017). In our study, we captured four different data streams with the wristband (i.e., HR, EDA, TEMP, BVP). We selected a state-of-the-art wristband device (Empatica E4), which has been measured up to the ECG gold standard (Greene et al., 2016), and data streams that have been used to infer various learning-related constructs in previous works (e.g., Di Lascio et al., 2017; Giannakos et al., 2019). Thus, although different methodological decisions might have had a slight impact on the results, we would not expect major deviations. Moreover, consideration of additional features (e.g., speaking periods, listening periods) may have offered additional insights and increased the accuracy of the results.

Finally, three different questionnaire-based attitudinal factors were selected to provide the ground truth (i.e., SAF, USE, and PER). These are validated scales, and their selection was grounded in the literature; however, it is arguable that different scales—such as the Course Experience Questionnaire (Ramsden, 1991), Course Flow Measure (Shin, 2006), Marzano Scale, PANAS, etc.—could had been used. In addition, it is known in HCI that the formulation of those scales induces certain bias (Müller et al., 2014), and sometimes the accuracy depends on various factors (e.g., context, participants' fatigue) (van Berkel et al., 2019b, 2018a); thus, despite our great efforts to minimize such biases, it is important to mention that these conditions might have had an impact on the results. A more frequent ranking of the experience might have yielded more accurate human-labeled data (Van Gog et al., 2012); however, this will have introduced periods of disengagement and disruptions. This is why we decided to repeat the relatively short (25-minute) learning activity three times, and perform leave-one-subject-out cross-validation instead of introducing more frequent ranking of the experience (via questionnaire). Despite these limitations, self-reports for labeling the data set have been shown to be one of the best methods for explicitly measuring students' learning experience (Fredricks and McColskey, 2012). Thus, although we followed an ecological, but also accurate, research design, we understand that other methodological decisions may have played an important role in the results. However, our methodology includes a robust set of data streams and questionnaire factors that are common to contemporary HCI and learning research.

6. Conclusion and ongoing work

Overall, our work shows that wearable sensing technologies hold the potential to capture learning experience intuitively and in an almost real-time manner. We provide evidence that leveraging wearable sensing can be a viable method to accurately track students' engagement during various learning experiences, thereby providing unique possibilities for evaluating various learning designs and making informed decisions. Therefore, the incorporation of wearable sensing enables: (1) students and teachers to monitor and reflect on learning processes and regulate different aspects of their learning/instruction, and (2) HCI and learning technology researchers to examine complex learning experiences in more agile and accurate ways. Our findings indicate that wearable sensing can accurately inform quantified-self technologies about students' learning experience for enhancing self-monitoring, evaluation, and metacognitive reflection in learning.

The contribution of this paper is twofold: (1) by conducting an in-the-wild study that collected wearable sensing data and information on students' learning (via standardized questions) during a class activity we quantify the capacity of wearable sensing technology to give accurate predictions of students' learning; and (2) we identify the time needed for wearable sensing technology to provide enough data to enable accurate predictions.

Our results show that it is feasible to use physiological data to obtain insights on students' learning experience; however, further work is needed to expand our understanding about the capacities of physiological sensing so as to accurately capture learning experience, as well as enhancing the quality of human-labeled data. Future work needs to utilize more temporal dynamics, since for this paper we used aggregated features, and thus did not fully consider the importance of time. Possible next steps may also focus on defining features at the group level using the wristband data, which would help in developing group-level quantified-self technologies to support collaborative feedback tools and learning. Future research should also collect data from different learning activities, on a larger scale, and use different and repeated surveys for data collection. Cross-validating and extending our findings would allow us to build generalized prediction models, as well as identify learning activities in which we can most accurately predict students' learning experience. Utilizing different physiological sensing devices (e.g., eye-tracking, cameras) would allow us to triangulate our findings, improve the accuracy of our predictions, and explore how different data streams can inform human-labeled data. In addition, we intend to investigate whether a plausible association exists between different learning activities (e.g., passive lectures, hands-on labs), different durations of the learning activity, different parts of the day/semester, and different student groups (e.g., age, experience, instinctive motivation). This will allow us to build an integrated understanding of the potential (as well as the limitations) of wearable sensing technology to understand students' learning experience.

Acknowledgement

The authors would like to thank all the participants of this study. This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. This work has received funding from the Norwegian Research Council under the project FUTURE LEARNING (number: 255129/H20) and Xdesign (290994/ F20).

Appendix A. Variable Importance

Importance of the top 10 features to predict the three dependent variables using the SVM with a polynomial kernel. Values are scaled between 0 and 100.

Satisfaction Feature	Importance	Performance Feature	Importance	Usefulness Feature	Importance
1st DF HR	100	1st ACC BVP	100	Median TEMP	100
SD BVP	94.6	1st ACC HR	95.2	Energy TEMP	97.7
Kurtosis BVP	84.5	2nd ACC HR	93.6	Max TEMP	88.6
Mean TEMP	82.3	2nd ACC BVP	92.1	Kurtosis TEMP	87.4

Mean EDA	78.3	3rd ACC HR	90.3	Min EDA	86.8
Skewness HR	78.1	SD TEMP	81.4	1st ACC HR	86.1
1st ACC HR	77.6	2nd ACC BVP	78.9	2nd ACC HR	82.8
Max TEMP	76.5	Energy EDA	78.3	1st DF TEMP	79.3
Median TEMP	74.9	1st DF EDA	78.2	Max BVP	77.2
1st DF EDA	72.9	3rd ACC BVP	77.4	2nd DF TEMP	70.2

HR = heart rate; **BVP** = blood volume pressure; **TEMP** = temperature; **EDA** = electrodermal activity; **SD** = standard deviation; **DF** = dominant frequency; **ACC** = auto correlation coefficient

Appendix B. Leave-one-subject-out vs. Leave-one-group-out

Comparing the results from the two different validation schemes: leave-one-subject-out and leave-one-group-out. In the first scheme, data from one participant was reserved for testing; in the second scheme, data from all participants from one group was reserved for testing.

This shows that the models are not overfitted using the training data in leave-one-subject-out.

Results are from a *t*-test on the testing RMSE.

Dependent Variable	SVM-polynomial	SVM-radial	Random Forest	SVM—Linear
Performance	$T = -0.01$	$T = -0.07$	$T = -0.01$	$T = -0.01$
	$P = 0.98$	$P = 0.94$	$P = 0.99$	$P = 0.99$
Satisfaction	$T = -0.01$	$T = -0.03$	$T = -0.10$	$T = -0.04$
	$P = 0.98$	$P = 0.98$	$P = 0.91$	$P = 0.96$
Usefulness	$T = -0.13$	$T = -0.01$	$T = -0.34$	$T = -0.12$
	$P = 0.89$	$P = 0.99$	$P = 0.73$	$P = 0.89$

Appendix C. Shapiro–Wilk Normality test

	Random Forest	SVM Poly.	SVM Rad.	SVM Lin.
PER	0.94 (0.35)	0.93 (0.43)	0.92 (0.16)	0.91 (0.37)
SAT	0.85 (0.18)	0.86 (0.35)	0.84 (0.39)	0.85 (0.39)
USE	0.88 (0.39)	0.90 (0.23)	0.85 (0.30)	0.86 (0.45)

References

Ahonen, L., Cowley, B.U., Hellas, A., Puolamäki, K., 2018. Biosignals reflect pair-dynamics in collaborative work: EDA and ECG study of pair-programming in a classroom environment. Sci. Rep. 8 (1), 3138.

Arnold, K.E., Karcher, B., Wright, C.V., McKay, J., 2017. Student empowerment, awareness, and self-regulation through a quantified-self student tool. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. ACM, pp. 526–527.

Aslan, S., Mete, S.E., Okur, E., Oktay, E., Alyuz, N., Genc, U.E., Stanhill, D., Esme, A.A., 2017. Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. Educ. Technol. 53–59.

Barreto, A., Zhai, J., Adjouadi, M., 2007. Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In: International Workshop on Human–Computer Interaction. Springer, Berlin, Heidelberg, pp. 29–38.

Beardsley, M., Hernández-Leo, D., Ramirez-Melendez, R., 2018. Seeking reproducibility: assessing a multimodal study of the testing effect. J. Comput. Assist. Learn. 34 (4), 378–386.

Bednarik, R., 2012. Expertise-dependent visual attention strategies develop over time during debugging with multiple code representations. Int. J. Human–Comput. Stud. 70 (2), 143–155.

Biggs, J.B., Tang, C., 2007. Teaching for Quality Learning At University, 3rd edn. McGraw-Hill Education, Maidenhead.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series analysis: Forecasting and Control. John Wiley & Sons.

Choe, E.K., Lee, N.B., Lee, B., Pratt, W., Kientz, J.A., 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 1143–1152.

Cowley, B., Ravaja, N., Heikura, T., 2013. Cardiovascular physiology predicts learning effects in a serious game activity. Comput. Educ. 60 (1), 299–309.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16 (3), 297–334.

Csikszentmihalyi, M., 1997. Flow and the Psychology of Discovery and Invention. HarperPerennial, New York.

Csíkszentmihályi, M., 2008. Flow: The Psychology of Optimal Experience. Harper Perennial, New York.

Dillenbourg, P., 2013. Design for classroom orchestration. Comput. Educ. 69, 485–492.

Dillenbourg, P., Jermann, P., 2010. Technology for classroom orchestration. In: Khine, M.S., Saleh, I.M. (Eds.), New Science of Learning: Cognition, Computers and Collaboration in Education. Springer, pp. 525–552.

Dillenbourg, P., Matuk, C., Tissenbaum, M., 2016. Real-time visualization of student activities to support classroom orchestration. In: ICLS'16: Proceedings of the 12th International Conference of the Learning Sciences. 2. pp. 1120–1127.

Di Lascio, E., Gashi, S., Krasic, D., Santini, S., 2017. In-classroom self-tracking for teachers and students: preliminary findings from a pilot study. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, pp. 865–870.

Di Lascio, E., Gashi, S., Santini, S., 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. Proc. ACM Interact. Mobile Wearable Ubiquit. Technol. 2 (3), 103.

Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., Specht, M., 2016. Learning pulse: using wearable biosensors and learning analytics to investigate and predict learning success in self-regulated learning. In: CrossLAK, pp. 34–39.

D'Mello, S., Dieterle, E., Duckworth, A., 2017. Advanced, analytic, automated (AAA) measurement of engagement during learning. Educ. Psychol. 52 (2), 104–123.

Empatica, Inc. (2017). Have you done comparative studies or validation on electrodermal activity sensor?[Online]. Available at: https://support.empatica.com/hc/en-us/articles/203005295-Have-you-done-comparative-studiesor-validation-on-Electrodermal-activity-sensor.

Empatica, Inc. (2017). Have you done comparative studies or validation of the heart rate obtained from the E3?[Online]. Available at: https://support.empatica.com/hc/en-us/articles/200293658-Have-you-done-comparative-studies-or-validation-of-the-heart-rate-obtained-from-the-E3-E4-.

Eynon, R., 2015. The quantified self for learning: critical questions for education. Learn. Media Technol. 40 (4), 407–411.

Filak, V.F., Sheldon, K.M., 2008. Teacher support, student motivation, student need satisfaction, and college teacher course evaluations: testing a sequential path model. Educ. Psychol. (Lond) 28 (6), 711–724.

Fornell, C., Larcker, D.F., 1981. Evaluating structural equation models with unobservable variables and measurement error. J. Market. Res. 39–50.

Fredricks, J.A., Filsecker, M., Lawson, M.A., 2016. Student engagement, context, and adjustment: addressing definitional, measurement, and methodological issues. Learn. Instr. 43, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002.

Fredricks, J.A., McColskey, W., 2012. The measurement of student engagement: a comparative analysis of various methods and student self-report instruments. Handbook of Research on Student Engagement. Springer, Boston, MA, pp. 763–782.

Garbarino, M., Lai, M., Bender, D., Picard, R.W., Tognetti, S., 2014. Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In: 2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH). IEEE, pp. 39–42.

Giannakos, M.N., Sampson, D.G., Kidziński, Ł., 2016. Introduction to smart learning analytics: foundations and developments in video-based learning. Smart Learn. Environ. 3 (1), 12.

Giannakos, M.N., Sharma, K., Pappas, I.O., Kostakos, V., Velloso, E., 2019. Multimodal

data as a means to understand the learning experience. Int. J. Inf. Manage. 48, 108–119.
Gray, J.A., Diloreto, M., 2016. The effects of student engagement, student satisfaction, and perceived learning in online learning environments. NCPEA Int. J. Educ. Leadership Preparat. 11 (1), 98–119.
Greene, S., Thapliyal, H., Caban-Holt, A., 2016. A survey of affective computing for stress detection: evaluating technologies in stress detection for better health. IEEE Consumer Electron. Mag. 5 (4), 44–56.
Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L., 2006. Multivariate Data Analysis 6 Pearson Prentice Hall, Upper Saddle River, NJ.
Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. 50. Sage, Los Angeles, CA, pp. 904–908.
Hassenzahl, M., 2010. Experience design: technology for all the right reasons. Synth. Lect. Human-Cent. Inf. 3 (1), 1–95.
Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., Alt, F., 2017. EngageMeter: a system for implicit audience engagement sensing using electroencephalography. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, pp. 5114–5119.
Henrie, C.R., Halverson, L.R., Graham, C.R., 2015. Measuring student engagement in technology-mediated learning: a review. Comput. Educ. 90, 36–53.
Hernandez, J., Riobo, I., Rozga, A., Abowd, G.D., Picard, R.W., 2014. Using electrodermal activity to recognize ease of engagement in children during social interactions. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, pp. 307–317.
Hussain, M.S., Calvo, R.A., Chen, F., 2013. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. Interact. Comput. 26 (3), 256–268.
Hwang, G.J., 2014. Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective. Smart Learn. Environ. 1 (1), 4.
Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J.H., Dey, A.K., 2012. Factors influencing quality of experience of commonly used mobile applications. IEEE Commun. Mag. 50 (4), 48–56.
Junokas, M.J., Lindgren, R., Kang, J., Morphew, J.W., 2018. Enhancing multimodal learning through personalized gesture recognition. J. Comput. Assist, Learn. 34 (4), 350–357.
Kaller, C.P., Rahm, B., Bolkenius, K., Unterrainer, J.M., 2009. Eye movements and visuospatial problem solving: identifying separable phases of complex cognition. Psychophysiology 46 (4), 818–830.
Kay, R.H., Knaack, L., 2009. Assessing learning, quality and engagement in learning objects: the learning object evaluation scale for students (LOES-S). Educ.Technol. Res. Dev. 57 (2), 147–168.
Kuh, G.D., 2001. Assessing what really matters to student learning inside the national survey of student engagement. Change 33 (3), 10–17.
Kuh, G.D., Kinzie, J., Buckley, J.A., Bridges, B.K., Hayek, J.C., 2011. Piecing Together the Student Success Puzzle: Research, Propositions, and Recommendations: ASHE Higher Education Report 116 John Wiley & Sons.
Kuvaas, B., 2006. Work performance, affective commitment, and work satisfaction: the roles of pay administration and pay level. J. Organ. Behav. 27, 365–385.
Lee, V.R., Drake, J.R., Thayne, J.L., 2016. Appropriating quantified self technologies to support elementary statistical teaching and learning. IEEE Trans. Learn. Technol. 9 (4), 354–365.
Liaw, S.S., Huang, H.M., 2013. Perceived satisfaction, perceived usefulness and interactive learning environments as predictors to self-regulation in e-learning environments. Comput. Educ. 60 (1), 14–24.
Lupton, D., 2014. Self-tracking cultures: towards a sociology of personal informatics. In: Proceedings of the 26th Australian Computer–Human Interaction Conference on Designing Futures: the Future of Design. ACM, pp. 77–86.
Mangaroska, K., Giannakos, M.N., 2018. Learning analytics for learning design: a systematic literature review of analytics-driven design to enhance learning. IEEE Trans. Learn. Technol..
Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K., 2017. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affect Comput 8 (1), 15–28.
Müller, H., Sedley, A., Ferrall-Nunge, E., 2014. Survey research in hci. Ways of Knowing in HCI. Springer, New York, NY, pp. 229–266.
Papavlasopoulou, S., Sharma, K., Giannakos, M.N., 2018. How do you feel about learning to code? investigating the effect of children's attitudes towards coding using eye-tracking. Int J Child Comput Interact 17, 50–60.
Pham, P., Wang, J., 2016. Adaptive review for mobile mooc learning via implicit physiological signal sensing. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, pp. 37–44.
Pijeira-Díaz, H.J., Drachsler, H., Järvelä, S., Kirschner, P.A., 2016. Investigating collaborative learning success with physiological coupling indices based on electrodermal activity. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM, pp. 64–73.
Pijeira-Díaz, H.J., Drachsler, H., Kirschner, P.A., Järvelä, S., 2018. Profiling sympathetic arousal in a physics course: how active are students? Journal of Computer Assisted Learning 34 (4), 397–408.
Prieto, L.P., Sharma, K., Kidzinski, Ł., Dillenbourg, P., 2017. Orchestration load indicators and patterns: in-the-wild studies using mobile eye-tracking. IEEE Transactions on Learning Technologies 11 (2), 216–229.
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.
Raaijmakers, S.F., Baars, M., Schaap, L., Paas, F., Van Gog, T., 2017. Effects of performance feedback valence on perceptions of invested mental effort. Learn Instr 51, 36–46.
Raca, M., Dillenbourg, P., 2014. Classroom social signal analysis. Journal of Learning Analytics 1 (3), 176–178.
Rahman, M.M., Chowdhury, M.A., Fattah, S.A., 2018. An efficient scheme for mental task classification utilizing reflection coefficients obtained from autocorrelation function of eeg signal. Brain Inf. 5 (1), 1–12.
Ramsden, P., 1991. A performance indicator of teaching quality in higher education: the course experience questionnaire. Studies Higher Education 16, 129–150.
Ross, S.M., Morrison, G.R., 2004. Experimental research methods. Handbook Res. Educ. Commun. Technol. 2, 1021–1043.
Ruiz, S., Charleer, S., Urretavizcaya, M., Klerkx, J., Fernández-Castro, I., Duval, E., 2016. Supporting learning by considering emotions: tracking and visualization a case study. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM, pp. 254–263.
Samuelsen, K.M., 2012. Part v commentary: possible new directions in the measurement of student engagement. Handbook of Research on Student Engagement. Springer, Boston, MA, pp. 805–811.
Sánchez, R.A., Huero, A.D., 2010. Motivational factors that influence the acceptance of Moodle using TAM. Comput. Human Behav. 26 (6), 1632–1640.
Schmidt, M., Tawfik, A., Earnshaw, Y., and Jahnke, I. (2019). Learner and user experience research: an introduction for the field of learning design & technology. Available at:https://edtechbooks.org/ux.
Segars, A.H., 1997. Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems research. Omega (Westport) 25 (1), 107–121.
Selwyn, N., 2015. Data entry: towards the critical study of digital data and education. Learn. Media Technol. 40 (1), 64–82.
Sharma, K., Papavlasopoulou, S., Giannakos, M., 2019a. Joint emotional state of children and perceived collaborative experience in coding activities. In: Proceedings of the 18th ACM International Conference on Interaction Design and Children. ACM, pp. 133–145.
Sharma, K., Papamitsiou, Z., Giannakos, M., 2019b. Building pipelines for educational data using AI and multimodal analytics: a "grey-box" approach. Brit. J. Educ. Technol..
Shin, N., 2006. Online learner's 'flow' experience: an empirical study. Brit. J. Educ. Technol. 37 (5), 705–720.
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. J. Neurosci. Methods 180 (2), 304–316.
Skuballa, I.T., Dammert, A., Renkl, A., 2018. Two kinds of meaningful multimedia learning: is cognitive activity alone as good as combined behavioral and cognitive activity? Learn, Instr, 54, 35–46.
Spector, J.M., 2014. Conceptualizing the emerging field of smart learning environments. Smart Learn, Environ, 1 (1), 2.
Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., Picard, R., 2015. Automatic identification of artifacts in electrodermal activity data. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 1934–1937.
Van Gog, T., Kirschner, F., Kester, L., Paas, F., 2012. Timing and frequency of mental effort measurement: evidence in favour of repeated measures. Appl, Cogn, Psychol, 26 (6), 833–839.
Wang, C., Cesar, P., 2015. Physiological measurement on students' engagement in a distributed learning environment. In: PhyCS, pp. 149–156.
Wang, R., Wang, W., daSilva, A., Huckins, J.F., Kelley, W.M., Heatherton, T.F., Campbell, A.T., 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2. pp. 43.
van Berkel, N., Budde, M., Wijenayake, S., Goncalves, J., 2018a. Improving accuracy in mobile human contributions: an overview. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, pp. 594–599.
van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., Kostakos, V., 2019a. Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, New York, NY.
van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., Kostakos, V., 2019b. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. Int, J, Hum, Comput, Stud, 125, 118–128.
van Berkel, N., Hosio, S., Goncalves, J., Wac, K., Kostakos, V., Cox, A., 2018b. MHC'18: international workshop on mobile human contributions: opportunities and challenges. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, pp. 590–593.
Van Gog, T., Paas, F., 2008. Instructional efficiency: revisiting the original construct in educational research. Educ, Psychol, 43, 16–26.
Van Zele, E., Vandaele, P., Botteldooren, D., Lenaerts, J., 2003. Implementation and evaluation of a course concept based on reusable learning objects. J. Educ. Comput. Res. 28 (4), 355–372.
Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R., 2014. The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans. Affect. Comput. 5 (1), 86–98.
Xu, R., Zhang, C., He, F., Zhao, X., Qi, H., Zhou, P., ..., Ming, D., 2018. How physical activities affect mental fatigue based on EEG energy, connectivity, and complexity. Front. Neurol. 9.