Towards Automatic Real-Time Estimation of Observed Learner’s Attention Using Psychophysiological and Affective Signals: The Touch-Typing Study Case

Marko Meža, Member, IEEE, Janja Košir, Gregor Strle, and Andrej Košir, Senior Member, IEEE

Abstract—This article presents an experimental study on the real-time estimation of observed learners’ attention given the task of touch-typing. The aim is to examine whether the observed attention estimates gathered from human raters can be computationally modeled in real time, based on the learner’s psychophysiological and affective signals. A key observation from this study is that the observed attention varies continuously and throughout the task. The findings show that a relatively high sampling interval is required for the modeling of observed learners’ attention, which is impossible to achieve with traditional assessment methods (e.g., between-session self-reports). The results show that multiple linear regression models were relatively successful at discriminating low and high levels of the observed attention. In the best case, the within-learner model performed with the goodness-of-fit adjusted $R^2_{adj} = 0.888$ and RMSE = 0.103 (range of the attention scores 1–5). However, the multiple linear model underperformed in the estimation of the observed attention between learners, indicating that the differences among the learners are often significant and cannot be overcome by a general linear model of attention. The between-learner model achieved an adjusted $R^2_{adj} = 0.227$ and RMSE = 0.708, explaining only 22.7% of the variability. The influence of individual psychophysiological and affective signals (eye gaze, pupil dilation, and valence and arousal) on the estimation of the observed attention was also examined. The results show that both affective dimensions (valence and arousal), as well as the EyePos2D offset (the distance of an eye from the average position in the x-y plane parallel to the screen), and the EyePos-Z (the distance of an eye from the screen) significantly and most frequently influence the performance of the within-learner model.

Index Terms—attention estimation, rater agreement, psychophysiological and behavioral signals

I. INTRODUCTION

ATTENTION is crucial for effective learning. It explains why and how we process certain information and block or ignore other information, and it gives insight into individual’s learning habits. Experienced human tutors are capable of intuitively modulating a learner’s attention toward a task in real time in order to achieve higher learning gains [1], [2]. In e-learning, where a human tutor is replaced by an intelligent tutoring system (ITS), continuously modeling and responding to a learner’s attention is much more difficult [3], [4]. Such a system would also need to be attention-aware and capable of estimating and modulating learners’ attention in order to improve their performance and engagement [5], [2].

There are several ways of estimating a learner’s attention during a learning process in ITS. For example, attention can be estimated based on feature extraction from eye gaze, facial emotion expressions, head posture, body posture, and gestures, among others [6], [7], [3], [8], [9], [10], [4]. The extracted features are subsequently used in machine learning models of attention and for determination of weak ground truth. Typically, the weak ground truth is determined based on learners’ self-reports of attention sampled either concurrently with the activity or retrospectively (e.g., from videos), or both [11], [3], [9], [12]. Additionally, human raters are often employed to retrospectively rate learners’ attention from video to establish weak ground truth. Attention estimation approaches based on self-reports are currently widely used in research studies, but they have several limitations. For one, self-reports are intrusive and require a learner to consciously switch between two tasks, thus interrupting their attention on the primary task. Furthermore, attention can only be sampled occasionally and for short periods of time, or retrospectively. Finally, given these limitations, the approaches are difficult to apply on a large scale or over extended periods of time.

One of the biggest challenges of ITSs is to develop mechanisms for automated real-time estimation of learners’ attention. Recent work by Whitehill et al. [13], D’Mello [4], Bixler et al. [14], and Hutt et al. [15] is an important step in this direction. The automated attention-aware ITSs pose several advantages over the aforementioned self-report approaches. They employ computational modeling of attention, rely on psychophysiological and behavioral signals for feature extraction, and can estimate and respond in real time without interrupting the learner. Moreover, the automated estimation of attention allows for continuous real-time intervention to redirect the learner’s attention to the task at hand, initiate particular strategy, or set an appropriate exercise or task level for optimal performance [14], [4]. The automated attention-aware ITSs should also be capable of detecting and modulating a learner’s affective state. In addition to attention, affective state is a key indicator of a learner’s engagement [16], [17]. Affect-aware ITSs are also effective at regulating negative emotions (such as anxiety.
frustration, boredom, and disengagement) that inevitably arise from learning in the ITS context [18], [11], [10]. Recent studies in affective computing and learning show that affect-aware ITSs are superior to traditional ITSs in terms of improved attention and learning [19], [5], [20], [8], [10].

This article presents an approach towards automatic real-time estimation of the attention of an observed learner while interacting with the ITS – an intelligent touch-typing tutor. The goal is to develop a robust and computationally efficient model that estimates the observed attention unobtrusively and in real time, requiring no additional input from the learner. The model estimates the observed attention from the learner’s psychophysiological and affective sensor signals. Human raters are employed to determine the initial weak ground truth of attention and to train the model.

An obvious advantage of the proposed approach is high sampling of the psychophysiological and behavioral signals. Unlike self-reports, wherein the learner’s concurrent rating of attention can only be obtained by lower the sampling intervals 1 (see [12], [4]), the presented approach estimates the dynamics of the learner’s attention continuously to detect potentially more fine-grained variations.

It is important to note that the presented attention estimation model is memoryless. It relies entirely on real-time sensory data from eye and gaze movement, pupil dilation, and affective dimensions of valence and arousal, thereby limiting a number of features that would require additional complexity and processing power. For example, it does not take advantage of the two features most typically observed in studies of attention utilizing eye-tracking – saccades and fixations [3]. The aim of the present study is to test the feasibility of such an approach and its potential for ‘in-the-wild’ applications, which generally require robustness and simplicity.

A. Overview

The article is structured as follows. Section II presents the related work on attention and learning in ITSs, as well as the computational approaches for the estimation and ground truth determination of attention. Section III presents the apparatus used in the experiment. It consists of two main components: a) an intelligent typing tutor equipped with psychophysiological and behavioral sensors, and b) the attention estimation tool. Section IV presents the experimental study, providing details about the experiment design, the operational definition of the observed attention, the attention annotation, the data preprocessing, the calculation of the observed attention estimate, and the evaluation procedure. Section V presents the results, evaluating the performances of several within-learner and between-learner attention estimation models. This is followed by the discussion in Section VI, with conclusions and direction for future work presented in Section VII.

1For example, Monkaresi et al. [12] argue that the sampling interval of concurrent self-reports should not be higher than 2 minutes to avoid interruptions that affect a learner’s attention and performance on a primary task.
independence rather than physiological independence” [41, p. 138]. For example, common brain areas of activation that mediate saccades and attention shifts point to common neural substrates [44], [45], [46]. Similarly, in a series of neuropsychological experiments, Katsuki and Constantinidis [47] showed that the top-down and bottom-up attentional processes share a common neural apparatus, the frontoparietal network.

The following sections discuss attention and learning in ITSs.

B. Attention in e-Learning

Observing students while learning is an important educational practice. Various techniques are used to help students understand the information relevant to a specific task and guide their attention in order to maximize their learning capabilities and outcomes [1]. In a traditional teaching environment, the aforementioned scenarios depend to a large extent on the teacher, their teaching methods, and the curriculum. In order to prevent a major loss of attention, a good teacher will strive to eliminate irrelevant stimuli from the learning environment and will employ different techniques and strategies (e.g., self-monitoring) to engage students and achieve intended goals.

In e-learning, a teacher’s role is replaced by an ITS. The procedural aspects related to learning and knowledge acquisition within an ITS need to be carefully defined [2]. Significant emphasis must be given to the human–computer interaction and design of user interfaces that encourage learning through user engagement and optimize (reduce) mental load for a given task with regard to attention [48], [49], [42], [50]. The latter is crucial. According to the studies of cognitive load and attention in e-learning [51], [52], [49], several aspects should be considered to reduce the effects of an extraneous cognitive load in the ITS. These are a) the worked example effect, wherein worked examples should be used to guide a user through problem states and provide concrete solutions to avoid excessive mental load; b) the split-attention effect which underlies the chunking of information and its presentation in an integrated way; c) the modality effect, which is an effect of extraneous load on the human cognitive system when too many modalities are activated at the same time; and d) the redundancy effect of repeating the same information through multiple sources [53], [52], [49]. Moreover, several studies argue that the ITS needs to be user- and affect-aware to successfully exploit the attention processes of its users [5], [11], [9], [10], [4]. As [48, p. 18] point out, “the findings about our limited attentional resources – and about how we rely on attentional signals in collaboratively – have significant implications for how we design computational systems and interfaces.” The following section briefly presents ITSs.

C. Intelligent Tutoring Systems

Several ITSs exist that support some level of user-awareness to amplify attention and overall engagement, ranging from emotion-aware to meta-cognitive. Two of the more relevant examples are Gaze Tutor [11] and Affective AutoTutor [54]. Gaze Tutor promotes learning and engagement based on the user’s affective state (boredom and disengagement) [11]. It uses eye-tracking software to monitor the user’s gaze and analyzes their affective state with the aim to boost their attention and guide them back to the task at hand. Based on the user’s affect, it then tries to re-engage and re-orientate the user’s attention to the important areas of the interface. The results showed that such approach was successful at reorienting users’ attention and also had a positive effect on learning gains [11]. Similarly, AutoTutor/Affective AutoTutor integrates both affective and cognitive modeling to support learning and engagement, tailored to the individual user [54]. Help Tutor [55] is a meta-cognitive variation of AutoTutor that aims to develop better general help-seeking strategies for students. Cognitive Tutor [56] is an instruction-based system for mathematics and computer science. The tutor integrates a problem-solving interface, a curriculum guided by a domain expert, a cognitive model for solving problems (in relation to the task at hand), and instruction and production rules [56]. Cognitive Tutor supports guided learning-by-doing, based on the self-explanation technique. The results of the study conducted by Aleven and Koedinger showed improved understanding for students using the self-explanation technique compared to students who did not explain steps [57]. MetaTutor [58] aims to model the complex nature of self-regulated learning and helps learners to develop self-regulated learning strategies. Many other ITSs exist, some as constraint-based intelligent tutoring systems that model instructional domains at an abstract level [59] or that model skills. For example, [60] developed a tutoring system for typing based on using a touch screen. The system detects and analyzes errors and suggests instructions to improve the user’s typing performance. In general, several studies on affective learning indicate the superiority of the affect-aware over the non-affect-aware ITSs in terms of improved attention and significant increases in performance in learning [57], [19], [61], [20], [8], [10]. The following sections present various approaches for measuring attention and determining its weak ground truth.

D. Measuring Attention from Behavioral Signals

There is a large amount of scientific literature related to measuring attention from behavioral signals, focusing predominantly on the non-verbal behavioral signals such as eye gaze, facial expression, body posture, and gesture interaction, among others [62], [63], [64], [7], [65], [4], [14]. An extensive overview of general conceptual and technical challenges related to modeling of attention in learning environments is presented in [66].

Eye gaze is one of the most important indicators of human attention, and several studies use eye tracking methodology to measure eye gaze [3]. Eye tracking is used both as an input modality and as a source of information on relevant user states and processes [62]. For example, it can be used to detect attention in order to increase learners’ bandwidth in an intelligent tutoring system [62], or to assess attention allocation when performing a task [66], [67], [3]. Koedinger and Anderson [68] used eye tracking to detect learners’ attention patterns on a computer educational game Prime Climb and studied the effects of personalized elicited and non-elicited hints during
the game. Asteriadis et al. [69] presented an un-intrusive system that evaluates the behavioral state of the learner (level of interest and engagement) in an e-learning environment. The system uses a web camera to measure the learner’s eye gaze and head position while they are reading a document on a computer screen. Farhan et al. [70] presented a methodology for measuring and analyzing visual engagement and visual attention for teachers and students in multimedia-based e-learning, based on eye tracking from video lectures. Another study of attention in e-learning coupled eye gaze with speech information to determine user’s interests by using a portable camera and microphone system to acquire human–computer interaction data [71]. User’s attention was then classified with a web-based learning system to estimate their attention level: attentive, not looking, and speaking.

An excellent review of the research on eye gaze and attention in ITS, discussing a range of attentional states (e.g., overt inattention, covert inattention, zone-outs, tune-outs, and focused, alternating, and divided attention), is presented in [4]. The author presents a study on attention-aware learning and argues for a three-component attentional computing architecture for measuring observed attention. A learner’s attention is measured using eye tracking, with integrated interventions to dynamically redirect attention toward the required area. In [14], the authors studied mind wandering – involuntary shifts of attention from the primary task – during reading. Eye gaze (fixation, saccade duration, and pupil dilation) and contextual cues were recorded during the task of reading using eye tracking, mind-wandering probes in the form of audio beeps, and students’ self-reports. The article also provides an excellent up-to-date review of attention estimation and mind-wandering detection methods (see also [15]).

In addition to eye gaze, facial expressions are also frequently used to evaluate human attention and emotional engagement while learning with ITSs [9], [10], [4]. Jacques et al. [9] used MetaTutor to explore the usefulness of eye tracking for predicting emotions relevant to learning, specifically boredom and curiosity, and to detect disengagement. Whitehill et al. [13] explored approaches for automatic recognition of student engagement from static images of facial expressions. The authors trained a binary classifier for the automatic discrimination of engagement from and image of a human face, based on the action units from The Facial Action Coding System (MLR CERT), among other features. The results showed that human engagement judgments reliably agree when discriminating low versus high degrees of engagement and that automated engagement detectors perform with accuracy comparable to that of humans. Monkaresi et al. [12] studied user engagement through non-obtrusive video-based observation of learner’s facial expressions and heart rate while the learner performed a writing task. The ground truth for the engagement levels was obtained through self-reports gathered concurrently (every 2 minutes) and retrospectively (one week after the experiment based on the students’ video recordings). Classifiers were then used to model engagement. Grafsggaard et al. studied the affective and cognitive processes in learning using facial emotion recognition [8]. The authors were able to predict frustration and learning outcomes from facial expressions early in the tutoring session. Their results show “three significant relationships between facial expression, frustration, and learning: 1) Action Unit 2 (outer brow raise) was negatively correlated with learning gain, 2) Action Unit 4 (brow lowering) was positively correlated with frustration, and 3) Action Unit 14 (mouth dimpling) was positively correlated with both frustration and learning gain” [8, pp. 160].

E. Weak Ground Truth Determination of Observed Attention

As discussed above, most ITSs rely on visual cues to measure and estimate learners’ attention. Additionally, human raters are often used to establish a weak ground truth (WGT) of observed learner’s attention while performing a certain task. For example, the study by [13], presented in the previous section, used human raters to establish the WGT for attention. Their results show it is possible to automatically recognize engagement from an image of a human face, with accuracy comparable to that of human raters. The authors found that human raters reliably agreed when discriminating low and high degrees of engagement (Cohen’s kappa = 0.96), however, the reliability decreased with finer discrimination (Cohen’s kappa = 0.56) [13]. In another study, the authors used a crowdsourced method (Amazon’s Mechanical Turk) to obtain human annotations of viewer’s affective responses to video [72]. The authors’ goal was to establish the WGT for predictive modeling of viewer’s affective responses (boredom), to be used in multimedia retrieval and recommendation systems. Crowdsourcing was used to gather a large number of annotations and close the gap between the raters and the target (viewer) group. The authors provide a list of recommended practices for gathering WGT using crowdsourcing techniques [72]. Other studies have used self-reports for ground truth determination. In Gaze Tutor, the WGT was established from participants’ self-reports on their level of engagement after each lecture. Participants were asked to rate their level of engagement (ranging from very bored to very engaged) for the beginning, middle, and end of each lecture on a six-point scale [11]. Additionally, the WGT was established based on participants’ self-reports on their perceived performance, user satisfaction, and task difficulty [11]. In MetaTutor, the WGT was also determined from learners’ self-reports. It was then used for the evaluation of several machine learning and feature selection techniques to predict students’ self-reported emotions from gaze data features [9].

1) Rater Agreement Coefficients: WGT is typically established with some type of rater agreement coefficient. The rater agreement coefficients describe the degree of agreement among raters for a particular observation, given a particular level of measurement (e.g., nominal, ordinal, interval, or ratio) and specific procedure. This procedure is, in most cases, simplified by cleaning up the rating data and then averaging the ratings from the previous step to obtain WGT by either mean or median operation. An extensive overview of the main statistical methods for measuring rater agreement is given by

The term ‘weak’ is used to denote that the phenomenon can only be observed and estimated indirectly.
Cohen’s kappa [76] and weighted kappa [77] measure the nominal scale agreement between two raters. An adaptation of Cohen’s kappa, Fleiss’ kappa [78], can measure the agreement among multiple raters. Krippendorff’s alpha [79] is widely used in content analysis and gives a measure of agreement achieved when coding a set of units of analysis in terms of the values of the variable. It can be used with nominal, ordinal, interval, and ratio levels of measurement, on any number of raters, and can account for small sample size [80]. Intraclass correlation coefficient (ICC) measures the proportion of variance of an observation due to between-subject variability in the true score [81]. ICC is a version of Krippendorff’s alpha wherein all raters evaluate all cases [79] and the agreement among the raters is then measured on a scale from 0 to 1, where the high values represent high agreement among raters. For the experimental study presented in this article, ICC was used to evaluate the agreement on attention ratings, as it takes into account both ratings for the individual segments and the correlation between raters [82].

The following sections present the experimental apparatus and the design of the study.

III. Experimental Apparatus

The experimental apparatus is comprised of two parts: the intelligent typing tutor equipped with typing lectures and sensors, and the attention annotation tool for rating the observed learner’s attention.

A. The Intelligent Typing Tutor

The intelligent typing tutor is a web-based application. It challenges the learners with a series of expert-designed touch-typing exercises [83], tracks their performances, records their psychophysiological and behavioral signals from the sensors (eye position, eye gaze, pupil dilation, facial emotion expressions), and sends the data to the server.

The graphical interface of the tutor is presented in Fig. 1. Each learner is shown a series of characters on the screen together with a series of related keyboard keys required as an input. The latter is presented as an image of both hands with the finger relevant for a particular character highlighted. The progress is tracked with a progress bar—a line encircling the in-focus character. The correct inputs are marked with characters colored in green and incorrect inputs are marked with characters colored in red. The performance also affects the smiley face, which becomes neutral upon a series of incorrect inputs. The small marker on the right edge of the typing tutor’s interface has no relevance for the learner—it links to a sensor debugging mode.

B. Psychophysiological Sensor Signals

Psychophysiological sensor signals are continuously recorded in real time during the typing sessions. They measure the following psychophysiological and affective signals: eye position, eye gaze, pupil dilation, and facial emotion expressions. The apparatus used for the measurement is the Tobii eye-tracking tool for analysis of eye gaze and Noldus FaceReader tool for facial emotion recognition, which are described in the following sections. Typing sessions are video-recorded for later human rating of learners’ attention. Connecting sensors and the video camera to a single computer (on which the typing tutor was running) simplified time synchronization issues, as all recorded time-stamps from different sensors used the same internal clock.

1) Capturing Eye Gaze: Tobii Eye Tracking: Tobii X2-30 eye-tracker4 is used to track learner’s gaze position, eye position, and pupil dilation. It is attached to the bottom of the monitor. Tobii eye tracker uses an infrared light source to generate reflection patterns on the corneas of the learner’s eyes. Reflection patterns are captured by the eye tracker’s image sensors and processed to obtain a 3D position of each eyeball, gaze direction, and pupil dilation. Signals are tracked separately for each eye.

The sampling frequency during the experiment was approximately 30 Hz. Each data sample consisted of a time-stamp (used later for synchronization), a 3D vector eye position, a 3D vector gaze position, and a scalar pupil diameter. Some additional signals were calculated from the measurements gathered by the Tobii eye tracker. The Euclidean distance of the gaze from the average gaze position and the Euclidean distance of the eye from the average eye position in the x-y plane (parallel to the screen) were calculated from each data sample and included in the dataset.

The following psychophysiological sensor signals (all for left eye) were used:

- **GazeOffset**, which is the calculated Euclidean distance of the gaze from the average gaze position,
- **EyePos2D-offset**, which is the calculated Euclidean distance of the eye from the average eye position in x-y plane (parallel to the screen)
- **EyePos-z**, which is the distance of the eye from the screen, reported directly by the eye tracker
- **PupilDilation**, which is the diameter of an eye’s pupil, reported directly by the eye tracker

![Fig. 1. The user interface of the intelligent typing tutor.](http://www.tobii.pro.com/siteassets/tobii-pro/user-manuals/tobii-pro-x2-30-eyetracker-user-manual.pdf)

---

3Each session follows an expert-designed exercise plan specified in the configuration file.

4http://www.tobii.pro.com/siteassets/tobii-pro/user-manuals/tobii-pro-x2-30-eyetracker-user-manual.pdf
2) Capturing Facial Emotion Expressions: Noldus FaceReader: Noldus FaceReader 6.1.5 is used for facial emotion recognition and detection of the learner’s behavioral signals (the affective state based on the dimensions of valence and arousal). FaceReader detects facial emotion expressions by analyzing 20 commonly used action units (AUs) [84]; for an extensive overview of FaceReader, see [85]. The FaceReader’s camera is positioned on the top of the monitor. The sampling frequency of the FaceReader during the experiment was approximately 6 Hz. Each data sample is time-stamped and includes the values for valence and arousal.

3) Recording of the Typing Sessions: Each session was video-recorded. The visual frame was set to include the learner’s upper body, hands and keyboard, as shown in Fig. 2. The recorded video resolution was set to 640x480 at 15 fps. The recorded video frame is time-stamped to allow for the subsequent alignment (time synchronization) of several data sources.

C. Attention Annotation Tool

The attention annotation tool is a web application that enables human raters to observe and rate the learner’s attention based on the video recordings of the typing sessions, as shown in Fig. 2. To familiarize the raters with their task, the annotation tool first displays the instructions about the rating procedure together with a preview of a video. The rating procedure goes as follows. Video playback is accompanied by an audio cue (through headphones) every 1.2 seconds, which serves as a temporal cue for the raters to provide their ratings.

The ratings are entered via the keyboard. The annotation tool records and time-stamps the ratings, along with the time-stamps for the predefined events (the timed audio cues). The annotation tool is used for facial emotion recognition and detection of the learner’s behavioral signals (the affective state based on the dimensions of valence and arousal). We further assume that the affective state of a learner plays a role in sustaining attention throughout the task [89], [90]. We are investigating overt, top-down, and goal-driven (intentional) and thus mostly endogeneous (voluntary eye movement) orienting of attention to a task at hand. Learners’ response times to a given task are subsequently measured to detect any spans of inattention or mind wandering (see Section VI-E). To summarize:

- we focus on the orienting of overt attention in relation to eye and gaze position, pupil dilation, and valence and arousal;
- we investigate the learner’s attentional focus – concentrating on a task at hand – as it is relatively simple to measure in terms of the learner’s performance (response times and errors) [91], [42];
- we model the observed attention estimate based on the human ratings of the observed learner’s attention (the raters’ observed learner’s eye movement, facial expressions, upper body posture, hand position, etc.);
- we don’t measure saccades and fixations or address covert shifts of attention in relation to eye movement.

Our main goal is to test the feasibility of the proposed approach, in part due to its potential for ‘in-the-wild’ applications rather than to subscribe to or test any existing theory of attention. From a feasibility perspective, and given the constraints of the experimental study, we aim to investigate the following questions: Can part of the observed attention be modeled in real time using limited sensory information from the eye, gaze, pupil dilation, and affective dimensions of valence and arousal? To what extent does the output of the model correlate with the WGT of the observed attention gathered by the human raters? And, to what extent do individual features contribute to the efficiency of the model?

1) Observed Attention Levels: To facilitate the rating, attention is categorized into five levels rated on a 5-degree Likert-scale:

- Very low
- Low
- Medium
- High
- Very high

The design of the computational model builds upon findings from the existing literature showing that eyes and gaze are a proxy for attention, as attention shifts and eye movements usually occur together [86], [43], [87], [88]. We further assume that the affective state of a learner plays a role in sustaining attention throughout the task [89], [90]. We are investigating overt, top-down, and goal-driven (intentional) and thus mostly endogeneous (voluntary eye movement) orienting of attention to a task at hand. Learners’ response times to a given task are subsequently measured to detect any spans of inattention or mind wandering (see Section VI-E). To summarize:

- we focus on the orienting of overt attention in relation to eye and gaze position, pupil dilation, and valence and arousal;
- we investigate the learner’s attentional focus – concentrating on a task at hand – as it is relatively simple to measure in terms of the learner’s performance (response times and errors) [91], [42];
- we model the observed attention estimate based on the human ratings of the observed learner’s attention (the raters’ observed learner’s eye movement, facial expressions, upper body posture, hand position, etc.);
- we don’t measure saccades and fixations or address covert shifts of attention in relation to eye movement.

Our main goal is to test the feasibility of the proposed approach, in part due to its potential for ‘in-the-wild’ applications rather than to subscribe to or test any existing theory of attention. From a feasibility perspective, and given the constraints of the experimental study, we aim to investigate the following questions: Can part of the observed attention be modeled in real time using limited sensory information from the eye, gaze, pupil dilation, and affective dimensions of valence and arousal? To what extent does the output of the model correlate with the WGT of the observed attention gathered by the human raters? And, to what extent do individual features contribute to the efficiency of the model?

1) Observed Attention Levels: To facilitate the rating, attention is categorized into five levels rated on a 5-degree Likert-scale:

- Very low
- Low
- Medium
- High
- Very high
type scale from 1 (very low) to 5 (very high). Each attention level is defined by a set of general attention cues (e.g., see [7]) – for example, ‘sitting still’ and ‘focused eye contact’ are representative of a high attention level. The observed attention levels are vaguely defined on purpose to allow for a certain degree of freedom, for example, for raters to observe any other potential behavioral signals relevant for estimating attention. Individual keyboard keys were assigned to each level, as shown in Fig. 2 and Table I.

B. Experimental Setup

The experimental study was conducted on fourth-year students (N = 32; 31 female and 1 male) from the Faculty of Education, University of Ljubljana. Participants answered a short survey including a self-report on their technique and experience with touch-typing. The survey showed the participants had no proper touch-typing skills – they exhibited different typing patterns, frequent visual feedback (keyboard to screen) while typing, and employed a maximum of 3–4 fingers on each hand while typing (the participants typically used index, middle, and ring fingers for typing, or any combination of the three). None of the participants had any previous experience with structured typing practice.

The typing sessions were conducted in an office equipped with a chair, a table, a computer keyboard and 17-inch laptop monitor, and sensor hardware. A mouse was used to set up the experiment. A video camera was placed on the tripod behind the table at an angle to capture the keyboard and the participant’s face, hands, and upper body posture, as shown in Fig. 2. The monitor was adjusted for each learner to be perpendicular to the eye-sight axis. The distance from the participant’s eyes to the screen was approximately 80 cm. Participants were allowed to adjust the keyboard and chair to a comfortable position. Small x-height character size displayed on the screen during typing was 2.5 cm. The lighting in the room was semi-controlled for the time of the experiment, with the outside natural lighting minimized (window blinds and curtains closed). The FaceReader requires a highly illuminated face, whereas the eye tracker illuminates the scene with its own IR light source. The office was lit by four 36 W fluorescent light tubes installed in light-diffusing luminaries on the ceiling, providing typical office illumination. The illuminance measured on a face was 90 lux. No additional artificial light sources were used.

The participants were first given an introduction to the study and the typing tutor application. They were instructed by the experiment coordinator and each followed the same experimental procedure. Next, the sensors were set up: the Tobii eye-tracking module was calibrated on each participant using a standard calibration procedure (requiring learners to look at designated points on the screen). Noldus FaceReader software was test-run and a video camera was set up. All participants were given a test training session, under the coordinator’s supervision, to familiarize themselves with the tutor’s interface and the touch-typing task. The duration of the test session lasted until the participants felt confident with the application (3–5 minutes on average). Before the start of the experiment, the coordinator left the office and the participants were left alone during the touch-typing session.

The typing session comprises a series of expertly designed touch-typing exercises (using all ten fingers) with various degrees of difficulty, adopted from a touch-typing course book [83]. The exercises are designed for Slovenian keyboard layout (qwertz layout with special characters for Slovenian language). The average duration of the typing session is 17 minutes, and each session is recorded with a video camera.

The learners are instructed to place their hands on the keyboard with both index fingers touching characters f and j (characters with tactile markings) – this is the initial hand position. Next, they are instructed to type in the characters that appear on the screen by using the designated fingers: the left little finger for a, the left ring finger for s, the left middle finger for d, the left index finger for f, etc. The correct finger for a particular key on the keyboard is highlighted in the stylized image of the hands in the tutor’s interface, directly below the character (see Fig. 1). The exercises follow the exercise plan with a series of tasks, each consisting of approximately 500 characters. These are presented to the learner in short intervals from various positions on the keyboard: from either a single row on the keyboard (e.g., “sladka ajda klala ladja...”), switching between the middle and the upper rows (e.g., “irfr juju rudar sraga kuhar...”), or between the middle and the lower rows (e.g., “jmm fvf jmjm lama ...”). All learners are given the same tasks in the same order (the characters that appear on the screen are not randomized) and the tasks are ordered by increasing difficulty to enable comparison among learners.

C. Attention Annotation

1) Gathering human ratings: Five human raters (three female, two male) were tasked with the annotation of the observed learner’s attention using the attention annotation tool (presented in Section III-C). The raters were fourth-year students from the same university as the participants, but they did not participate in the touch-typing sessions. The raters were instructed about the annotation procedure and the observed attention levels (see Section IV-A1), while also given the flexibility to rely on their own interpretation. Next, they were introduced to the attention annotation tool (see Section III-C) and encouraged to practice with it to familiarize themselves with the annotation task.

For the annotation, 15 participants (14 female, 1 male) were randomly selected from the typing sessions. A test segment with a duration of 6.5 minutes was selected from the overall session (17 minutes), starting at minute 4:30 and ending at minute 11:00 of the session. The attention annotation was conducted over two days, with each rater rating the attention for all 15 participants. The ratings were then computed in terms of their intra-class correlation coefficient ICC and used for the estimation of the observed attention (see Section IV-E).

8According to a recent study, "treating ratings as ordinal, instead of nominal values, generates less biased datasets and, in turn, more reliable models..." [92, p. 314].
The observed attention estimates are calculated using the truncated mean. For each learner, we observe \( n = 325 \) sampled attention estimation values given by the raters \( A_i \), where \( i = 1, \ldots, n \). Let us define \( \gamma \) as the constant indicating the percentage of omitted minimum and maximum values \( 0 \leq \gamma < 0.5 \), and let us assume \( g = \lfloor \gamma N_A \rfloor \), where \( \lfloor \cdot \rfloor \) stands for rounding down to the nearest integer, leading to same number of minimum and maximum values to be omitted. Let us define \( \text{sort}(\cdot) \) as a sorting operation. Using the \( \gamma \)-truncated mean definition, we describe the truncated mean of the observed attention for the \( k \)-th learner \( a_k^{\text{OAT}} \) as

\[
a_k^{\text{OAT}} = \left[ a_{k,1,1}^{\text{OAT}}, \ldots, a_{k,1,n}^{\text{OAT}} \right].
\]

where

\[
a_k^{\text{OAT}} = \frac{1}{N_A - 2g} \sum_{i=g}^{N_A-g} \text{sort}(a_{k,i,m}),
\]

Using (3), we take all \( m \)-th scores from the raters, sort them by size, omit minimum and maximum \( g \) ratings, and compute the mean value of the remaining ones. For example, for the 20% truncated mean on five observations, one may take \( g = \lfloor 0.20 \cdot 5 \rfloor = 1 \); hence, the truncated mean would be represented by a mean value with the minimum and the maximum value omitted. We sample \( n \) truncated means for each learner \( L_k \), obtained from \( N_A \) raters. After applying the ICC-based threshold, these values are used in the calculation of the observed attention estimate \( a_k^{\text{OAE}} \) (see Section IV-E).
2) Intra-class Correlation ICC as a Measure of Rater Agreement: Measuring the rater agreement for attention is a challenging task. One of the challenges is the idiosyncratic nature of human behavior and personal traits, which hinders attempts towards a more uniform modeling of a learner’s attention [42], [39]. Learners’ attention may vary significantly throughout a learning session and among the learners, as was also the case in this study. Another challenge is the decline of learners’ attention over prolonged periods of time [97]; however, this problem was not detected as the experimental period was relatively short. Moreover, human raters inevitably introduce additional bias in their estimation of attention.

To alleviate the latter problem, the intra-class correlation ICC was used as a measure of the agreement among raters. ICC was used for its ability to measure the within- and between-rater variability on the ordinal scale and the agreement among any number of raters [82].

ICC is defined by the formula

\[
ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2},
\]

where \(\sigma_B^2\) denotes the between-rater variance and \(\sigma_W^2\) denotes the within-rater variance [98].

The ICC values below 0.4 indicate a poor level of agreement, values between 0.4 and 0.59 indicate a fair level of agreement, values between 0.60 and 0.74 indicate a good level of agreement, and values above 0.75 indicate an excellent level of agreement among the raters [99].

Since the observed learner’s attention varies through time, often significantly, we estimate ICC for the time subintervals \(I_m\), \(m = 1, \ldots, n\), of length \(|I_m| = 40\) samples, as a sliding window obtaining ICC\(_{k,m}\) for each learner \(L_k\). The estimated ICC\(_{k,m}\) is assigned to the middle of the sliding window’s time. No weighting is applied. The length of the time subinterval \(I_m\) is a compromise between the competing demands of preserving the temporal dynamics and the optimal estimation accuracy. An example of such time-varying rater agreement is presented in Fig. 3.

3) The Rater Agreement Threshold: In order to obtain the relevant results from the human raters, it is essential that the ICC agreement among them is adequate. Relevant estimates of the observed learner’s attention \(a^{OAE}\) can be determined from the time segments with adequate ICC values:

\[
a^{OAE}_{k,m} = \begin{cases} 
a^{OAE}_{k,m}, & \text{if } ICC_{k,m} \geq ICC_T; \\
\text{undefined}, & \text{if } ICC_{k,m} < ICC_T. 
\end{cases}
\]

The calculation of \(a^{OAE}\) is therefore constrained by the inter-class correlation threshold ICC\(_T\), which removes the attention estimates with poor rater agreement. The threshold level depends on the ICC values defining the level of agreement. In the presented study, ICC\(_T\) is set to 0.4.

Table II shows the ICC \(\geq 0.4\) proportions and averages per learner.

An example of the observed attention estimate \(a^{OAE}\) calculation is illustrated in Figure 4, taken from a short interval (30 seconds) for a selected learner.
The following section presents the evaluation procedure and the design of the computational model used in the real-time observed attention estimation.

F. Evaluation Procedure

The following section describes the design of the computational model for estimating $\alpha^{OAE}$. The data was first preprocessed to reduce the noise and then resampled, as described in Section IV-D. This procedure left 325 unique data samples per learner. Using ICC$_T$ (ICC $\geq$ ICC$_T$, ICC$_T$ = 0.4), the number of available samples was reduced, as shown in Table II.

Each data sample is composed of one dependent variable $\alpha^{OAE}$ and six independent variables from the psychophysiological sensor signals. The latter include: the 2D eye position offset from the average eye position in the x-y plane parallel to the screen (EyePos2D-offset), the eye distance from the screen (EyePos-z), the gaze offset from the average gaze position (GazeOffset), and the pupil dilation (PupilDilation), as well as the affective dimensions of valence (Valence) and arousal (Arousal).

Multiple linear regression was used to model $\alpha^{OAE}$, using the ordinary least squares training algorithm. The models make predictions without prior history. The decision to model the observed attention with a linear model was based on the desire for simplicity and robustness, as well as on the preliminary analysis, which showed similar performance in comparison to non-linear alternatives (for a comparison of both models, see Section VI-D).

Several multiple linear regression models were trained for the within-learner (see Section V-A) and the between-learner observed attention estimation (see Section V-B). Their performance is reported in terms of the root mean square error (RMSE) and goodness-of-fit adjusted $R^2$. The influence of the individual psychophysiological and behavioral sensor signals on the model’s estimation of $\alpha^{OAE}$ is reported with the null hypotheses $H_0 = [\beta_{s_i} = 0]$ for each sensor signal $s_i$, where $\beta_{s_i}$ is an adjusted regression coefficient of the fitted multiple linear regression model for the individual sensor signal. The p-values ($p < \alpha = 0.05$) indicate the influence of the sensor signals on the model. To estimate the relevance of each signal across the learners, we report on the fraction of the learners for which a given signal influenced the model. We also report standardized multiple linear regression coefficients $\tilde{\beta}$, which also indicate the importance of the signal.

The following validation procedure was used. In the cases where the dataset for the training and validation is the same (e.g., for the within-learner modeling), ten-times-repeated random sub-sampling validation was used, with a randomly selected 67% of the data samples used in the training and the remaining 33% used in testing the model. The values obtained during the repeated training/testing runs were averaged across repetitions. In other cases, and depending on the task (see Section V-B), the model was either trained on one learner and tested on the others, or trained on all learners except the learner being tested. The same data was never used to train and test the model.

V. RESULTS

The following sections present the performances of the multiple linear regression models for the within-learner and the between-learner estimation of $\alpha^{OAE}$.

A. Within-Learner Model Performance

The multiple linear regression models for the within-learner modeling of $\alpha^{OAE}$ were created for each learner separately and evaluated using random sub-sampling validation, repeated ten times with a train/test ratio of 67/33. The data sets used in the modeling were created from the samples of $\alpha^{OAE}$ (dependent variable) and the sensor signals (independent variables). Only samples with ICC $\geq$ 0.4 were used (see Section IV-D).

Table III shows the within-learner model performances in terms of RMSE and adjusted $R^2$ per learner and for the whole test session. Learners 7 and 13 were deliberately selected in the presentation to illustrate the worst performing cases. The models make predictions without prior history. The scatter plots and time graphs show the observed vs. modeled $\alpha^{OAE}$s. The empty slots in the time graphs indicate the time intervals in which the ICC was below the threshold (ICC$_T$ < 0.4) and thus excluded from the evaluation.

The scatter plots and the RMSE distributions in Table III show that the models, in general, better predict the high and low levels of the $\alpha^{OAE}$s than the medium levels. This was also corroborated by the calculation of the overall error distributions across $\alpha^{OAE}$s for all models, presented in Fig. 5.

The influence of the individual sensor signals on the estimation of $\alpha^{OAE}$ was also analyzed. We report the null hypothesis $H_0 = [\beta_{s_i} = 0]$ for each sensor signal $s_i$, where $\beta_{s_i}$ is an adjusted regression coefficient of the fitted multiple linear regression models for the individual sensor signal. The p-values ($p < \alpha = 0.05$) shown in Table IV indicate the influences of the individual sensor signals on the model. Table IV also shows the standardized multiple linear regression coefficients $\tilde{\beta}$, which further indicate the influence of each signal on the model. To estimate the relevance of each signal across the learners, we report on the fraction of the learners for which a given signal influenced the model.

As shown in Table IV, among the features that significantly influenced the models’ predictions are the affective dimensions of valence and arousal, EyePos2D-offset, EyePos-z, and PupilDilation.
### TABLE III

The performance of the within-learner multiple linear regression models, for the five learners and the whole test sessions. The results are shown in terms of RMSEs and $R^2_{adj}$. The scatter plots and the time dependent graphs show the performance in terms of the observed (by human raters) vs. modeled aOAEs: on the X- and Y-axis respectively for the scatter plots, and on the time graphs, as the blue (observed) vs. red (modeled) lines representing the whole test session.

<table>
<thead>
<tr>
<th>Learner</th>
<th>RMSE</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.106</td>
<td>0.888</td>
</tr>
<tr>
<td>2</td>
<td>0.212</td>
<td>0.625</td>
</tr>
<tr>
<td>3</td>
<td>0.161</td>
<td>0.685</td>
</tr>
<tr>
<td>7</td>
<td>0.434</td>
<td>0.380</td>
</tr>
<tr>
<td>13</td>
<td>0.106</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Scatter plot | Time graph
---|---
![Scatter plot](scatter-plot-1.png) | ![Time graph](time-graph-1.png)

![Scatter plot](scatter-plot-2.png) | ![Time graph](time-graph-2.png)

![Scatter plot](scatter-plot-3.png) | ![Time graph](time-graph-3.png)

![Scatter plot](scatter-plot-7.png) | ![Time graph](time-graph-7.png)

![Scatter plot](scatter-plot-13.png) | ![Time graph](time-graph-13.png)
In general, the models’ performances vary among the learners, as shown in Tables III and IV. The model trained on learner 1 shows the best performance in terms of the adjusted $R^2_{adj}$ values. Overall, the models built on learner 1 performed best. Notable performances include the models built on learners 2, 3, 5, 9, and 10, whereas the model for learner 13 performed the worst ($R^2_{adj} = 0.001$). However, the scatterplot in Table III shows that the average $\alpha_{\text{OAE}}$ for learner 13 is around 3.5; therefore, even with a relatively small error (RMSE = 0.106), the adjusted $R^2_{adj}$ is nevertheless low.

### B. Between-Learner Model Performance

This section focuses on model generalizability – the between-learner model performance – with the following three tests. The aim of Test 1 is to evaluate how the model trained on the individual learner performs (in terms of $\alpha_{\text{OAE}}$) on the other learners. The aim of Test 2 is to evaluate how the models trained on all learners, except the one being tested, perform on each tested learner. The aim of Test 3 is to evaluate the performance of the general between-learner model trained on all learners.

1) **Test 1:** In the first test, separate multiple linear regression models were used for the between-learner evaluation of $\alpha_{\text{OAE}}$. Each model, previously used for the within-learner evaluation of $\alpha_{\text{OAE}}$, was tested on the rest of the learners (trained on one, tested on the others). The aim was to evaluate the generalizability of the individual within-learner model in terms of its performance across the other learners. Table V shows the models with relatively good performance. For example, the model trained on learner 12 was successful at modeling $\alpha_{\text{OAE}}$ for learners 1, 8, and 14, whereas the model trained on learner 14 was successful at modeling learners 9 and 1.

2) **Test 2:** The goal of the second test was to evaluate how the between-learner models generalize across learners. The models were trained on all learners except the learner being tested, and this procedure was then repeated for the remaining learners. The results are shown in Table VI with the reported RMSE and adjusted $R^2_{adj}$. The table also shows the influence of the individual signals on the model, with the p-values and the standardized multiple linear regression coefficients $\beta$.

It can be seen that this between-learner model performed poorly, achieving moderate results only for learners 3 and 9. This suggests that the differences among the learners are larger than the similarities conveyed by the model and that there is no one-size-fits-all solution (see Section VI for further discussion).

3) **Test 3:** In the third test, the general between-learner model was trained and tested on all learners using random sub-sampling, repeated ten times, with a train/test ratio of 67/33. The goal was to obtain a general estimation of the...
observed attention among the learners. As shown in Table VII, the reported goodness of fit is $R^2_{adj} = 0.227$, which is low, explaining 22% of the variability, with RMSE = 0.708 (in the range 1–5). The influence of the individual sensor signals on the model’s estimation of the between-learner $\beta^{\text{adj}}$ is reported by the p-values and the standardized multiple linear regression coefficients $\beta$.

VI. DISCUSSION

A. Advantages and Limitations of the Proposed Approach

The aim of the study was to test the feasibility of the memoryless model for the estimation of the observed learner’s attention by focusing only on the psychophysiological and behavioral signals that can be measured unobtrusively and in real time. This was done to test the feasibility of such an approach and its potential for ‘in-the-wild’ applications, which generally require robustness and simplicity.

The presented approach has obvious limitations as it over-simplifies several aspects of the bottom-up and top-down processing and attentional dis/engagement reported in the literature [100], [41], [46], [35]. For example, the presented computational modeling is memoryless and as such ignores two aspects most typically observed in the eye-tracking studies of attention: saccades and fixations [3]. Furthermore, it cannot detect covert phases of learners’ in/attention – learners can be in different attentional states than those expressed and measured by the sensor signals – nor can it effectively isolate goal-driven from stimulus-driven processes or gain any deep insight into their workings (see [41], [35] for a discussion on eye movement and orienting of attention). However, later analysis of the data did show that the potential spans of inattention and mind wandering didn’t significantly influence the models, as demonstrated by the high correlation between the response times and the WGT (see Section VI-E).

We have reservations with regard to more structured approaches for analyzing learners’ in/attention and mind wandering, as existing methods are overly invasive. For example, in their study on the automatic detection of mind wandering during reading, Bixler and D’Mello [14] used audio probes (in the form of beeps) as a cue for the participants to report their mind-wandering state while working on a task. The authors claim that such an approach gave them insight into the reader’s mind wandering. However, in our opinion, such an approach disrupts the reader and interferes with their task at hand, especially in cases where more frequent sampling is required. It is also questionable whether such an approach, which demands conscious reflection on the individual’s own attentional state, can be reliable or valid.

More general limitations of the presented study are the following. The models were designed to estimate the learner’s attention in the task of learning touch-typing. Cross-domain applicability of the models was not tested, nor were the model’s performances in different sessions and for the same learner. However, we believe that the presented approach using sensory signals for the estimation of the observed attention is applicable in other relevant domains with similar experimental setups.

The major advantages of the proposed approach are as follows:

- the model is capable of estimating the observed attention in real time, based on the psychophysiological and affective sensor signals;
- the attention estimates can be obtained with high sampling to account for dynamic changes in attention levels, as observed by human raters. Related studies report significantly lower sampling intervals (2 minutes) (e.g., [12]);
- the presented approach is relatively unobtrusive. It does not disrupt the learner with concurrent self-reports, another commonly used approach (e.g., see, [12], [15]).

The results show that the multiple linear regression models were relatively successful at discriminating the low and
high levels of the observed learner’s attention. The models performed relatively well for the within-learner attention estimation, barring the modeling of subtle or abrupt changes in the learner’s attention. The influence of the individual sensor signals on the estimation of observed attention was also examined. The results show that the affective dimensions of valence and arousal, the EyePos2D offset (the distance of an eye from the average position in the x-y plane parallel to the screen), and the EyePos-Z (the distance of an eye from the screen) most frequently and significantly influence the (within-learner and between-learner) models’ performances. Note that the between-learner models underperformed in their estimation of attention across learners, indicating that the differences among learners are often significant and cannot be overcome by a general linear model of attention. In our opinion, this is largely due to personality and behavioral differences among the learners that influence different psychophysiological reactions, various levels of emotion expressivity, and levels of engagement, as well as differences in working memory capacity and attention control, among other factors [101], [102], [37], [39].

B. Differences Among Learners

Table II shows the differences among the learners in terms of the available data samples, given ICC ≥ 0.4. In some cases, the proportion of ICC ≥ 0.4 samples is relatively high (68% for learner 5), whereas in others it is relatively low (9.5% 11.4% 12.6%, and 12% for learners 4, 8, 9, and 10). This shows how a learner’s personality and behavior (eye movement, gaze, pupil dilation, body posture, and emotion expressivity) can influence the measurements. Different behavior and emotion expressivity among the learners can also be observed in the video recordings of the sessions.

Often, these differences are significant. Fig. 6 compares the emotion expressivity of two learners (5 and 9) along the valence (x-axis) and the arousal (y-axis) dimensions for the whole test session. As shown in Fig. 6, the learners share positive arousal; however, learner 5 has significantly higher emotion expressivity compared to learner 9.

C. Classification Instead of Regression

The performance of the between-learner models was further evaluated via an alternative approach. A logistic regression classifier was trained on the observed attention data on two classes – low ($a_{\text{OAE}} < 2.5$) vs. high ($a_{\text{OAE}} ≥ 2.5$) level of attention – on the data for all learners with rater agreement ICC ≥ 0.4. The classifier successfully differentiated between the low and high levels of observed attention, with area under the receiver operating characteristic (ROC) curve (AUC) = 0.693. These results are comparable to those in the study on classification of mind wandering (using eye tracking and facial features) by Monkaresi et al. [12], who reported AUC = 0.758.

D. Performance Comparison of the Linear vs. Non-linear Models

In a preliminary study, the multiple linear regression model was compared with a non-linear model, multi-layer perceptron regressor, to evaluate the suitability of a linear model for modeling $a_{\text{OAE}}$. Both models were trained and tested separately for each learner, with the randomized shuffling validation 67/33 repeated ten times. The results in Table VIII show similar performances for both models in terms of RMSE (Spearman’s $\rho = 0.93$) and adjusted $R^2$ (Spearman’s $\rho = 0.9$), supporting the choice of the multiple linear regression model. In the study, the fitting algorithms and the goodness-of-fit $R^2$ estimation provided by the multiple linear regression model proved to be effective and stable.

E. Correlation Between the Response Times and the Observed Attention Estimates as a Proxy of Learner’s Attention

We investigated whether $a_{\text{OAE}}$ correlates with learners’ performance on the task of touch-typing. The evaluation was performed by analyzing the learner’s response time, that is, the delay between the displayed character and the response of pressing the character on the keyboard. The reasoning behind this is that the learner’s response times are influenced by their level of attention: higher attention yields quicker response times, and conversely, unusually long response times are an indication of a learner’s inattention and/or mind wandering. Results in Table IX show that there exists a moderate to strong negative correlation between the response time (the delay, hence the negative values) and $a_{\text{OAE}}$. The Spearman’s rank correlation coefficient is used for the data, which is not normally distributed.
An important observation taken from this study is that the observed learner’s attention varies continuously and throughout the tutoring session. To appropriately model these dynamic changes in attention, a relatively high sampling rate is required.

TABLE VIII
Performance comparison: the multiple linear regression model vs. the non-linear multi-layer perceptron regressor.

<table>
<thead>
<tr>
<th>Learner</th>
<th>RMSE Linear</th>
<th>RMSE Perceptron</th>
<th>( R^2_{adj} ) Linear</th>
<th>( R^2_{adj} ) Perceptron</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.106</td>
<td>0.114</td>
<td>0.887</td>
<td>0.879</td>
</tr>
<tr>
<td>2</td>
<td>0.217</td>
<td>0.215</td>
<td>0.614</td>
<td>0.619</td>
</tr>
<tr>
<td>3</td>
<td>0.161</td>
<td>0.158</td>
<td>0.676</td>
<td>0.681</td>
</tr>
<tr>
<td>4</td>
<td>0.101</td>
<td>0.077</td>
<td>0.462</td>
<td>0.554</td>
</tr>
<tr>
<td>5</td>
<td>0.309</td>
<td>0.325</td>
<td>0.635</td>
<td>0.622</td>
</tr>
<tr>
<td>6</td>
<td>0.144</td>
<td>0.189</td>
<td>0.401</td>
<td>0.232</td>
</tr>
<tr>
<td>7</td>
<td>0.428</td>
<td>0.481</td>
<td>0.380</td>
<td>0.307</td>
</tr>
<tr>
<td>8</td>
<td>1.118</td>
<td>0.554</td>
<td>0.323</td>
<td>0.219</td>
</tr>
<tr>
<td>9</td>
<td>0.206</td>
<td>0.288</td>
<td>0.819</td>
<td>0.749</td>
</tr>
<tr>
<td>10</td>
<td>0.177</td>
<td>0.162</td>
<td>0.782</td>
<td>0.798</td>
</tr>
<tr>
<td>11</td>
<td>0.230</td>
<td>0.846</td>
<td>0.464</td>
<td>0.182</td>
</tr>
<tr>
<td>12</td>
<td>0.078</td>
<td>0.099</td>
<td>0.300</td>
<td>0.160</td>
</tr>
<tr>
<td>13</td>
<td>0.124</td>
<td>0.111</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>14</td>
<td>0.385</td>
<td>0.308</td>
<td>0.087</td>
<td>0.124</td>
</tr>
<tr>
<td>15</td>
<td>0.203</td>
<td>0.187</td>
<td>0.266</td>
<td>0.318</td>
</tr>
</tbody>
</table>

TABLE IX
Correlation between the response times and \( \rho_{\text{OAE}} \). In most cases, the correlation coefficient shows moderate to strong linear correlation between \( \rho_{\text{OAE}} \) and the learner’s response time (keyboard delay). Shorter response times indicate higher attention.

<table>
<thead>
<tr>
<th>Learner</th>
<th>Spearman ( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.440</td>
</tr>
<tr>
<td>2</td>
<td>-0.243</td>
</tr>
<tr>
<td>3</td>
<td>-0.371</td>
</tr>
<tr>
<td>4</td>
<td>-0.338</td>
</tr>
<tr>
<td>5</td>
<td>-0.372</td>
</tr>
<tr>
<td>6</td>
<td>-0.218</td>
</tr>
<tr>
<td>7</td>
<td>-0.024</td>
</tr>
<tr>
<td>8</td>
<td>-0.586</td>
</tr>
<tr>
<td>9</td>
<td>-0.466</td>
</tr>
<tr>
<td>10</td>
<td>-0.570</td>
</tr>
<tr>
<td>11</td>
<td>-0.259</td>
</tr>
<tr>
<td>12</td>
<td>-0.151</td>
</tr>
<tr>
<td>13</td>
<td>0.059</td>
</tr>
<tr>
<td>14</td>
<td>-0.358</td>
</tr>
<tr>
<td>15</td>
<td>-0.184</td>
</tr>
</tbody>
</table>

VII. CONCLUSIONS AND FUTURE WORK
An important observation taken from this study is that the observed learner’s attention varies continuously and throughout the tutoring session. To appropriately model these dynamic changes in attention, a relatively high sampling rate is required (in the presented study, the sampling interval was 1.2 seconds). This cannot be achieved with traditional assessment methods (such as concurrent self-reports) or by using human raters to retrospectively estimate learners’ attention. As noted in the introduction to the article, such approaches have several limitations, specifically intrusiveness and the low sampling of the data, in addition to requiring learners to consciously switch between tasks, interrupting their attention on the primary task. These approaches are also difficult to apply on a larger scale or over extended periods of time.

The experimental study confirmed the usefulness of the selected psychophysiological and behavioral sensors for the within-learner estimation of attention. From the perspective of the intelligent tutoring system, and human–computer interaction in general, the dynamics of the interaction is extremely important. The proposed approach captures the dynamics of the observed attention, accounts for the rater agreement, and makes stable predictions in the attention estimation for individual learners.

The presented approach is a step towards an automated estimation of observed attention in real time. It employs psychophysiological and behavioral sensors that make estimation relatively unobtrusive and enable high sampling intervals. Our future work will focus on extending the presented study on a larger sample of learners. The inclusion of additional psychophysiological signal modalities, such as galvanic skin response, posture detection, 3D accelerometer, and gyroscope, is also planned to evaluate their potential for modeling observed attention. We also intend to repeat the experiment in related domains to evaluate the universality of the proposed models. Additionally, future work will investigate more advanced methods, taking into account various behavioral signals typical of a specific group of learners. The results from Table V suggest that there are different clusters of learners for which a particular within-learner model performs better. This indicates that several learners share some common behavior expressed by one particular model but not by others.

DISCLAIMER
This study was approved by the National Ethical Committee of the Faculty of Education, University of Ljubljana. All the procedures performed in the study are in accordance with the 1964 Helsinki declaration and its later amendments, APA Ethical Principles and Code of Conduct in psychological research, and institutional ethical rules for conducting research with human participants. Prior to their participation, the participants were informed about the experiment and the data being collected, and informed consent was obtained from all the participants.

REFERENCES


Marko Meža PhD, b.s.c. received his b.s.c. and Ph. D. degrees in Electrical Engineering from the University of Ljubljana in years 2001 and 2007. Since 2013 he is Assistant Professor at Faculty of Electrical Engineering, University of Ljubljana and a research member of User-adapted communications and ambient intelligence lab. His research interests cover medical and social signal processing using machine learning and datamining.

Janja Košir received the B. E. and M.E. in pedagogical science from the University in Ljubljana in 2005 and 2016, respectively. She is currently a teaching assistant of Didactic of special and rehabilitation pedagogy in the department of Special and Rehabilitation Pedagogy of the Faculty of Education of the University of Ljubljana. Her research interest among others is in assessing learning processes using the Intelligent Information and Communication Technologies (ICT) supported learning in education of people with special needs.

Gregor Strle received his B.A. degree from Philosophy and M.A. degree from Information Science from University of Ljubljana in 2002 and 2008, respectively, and his PhD in Cognitive Sciences from University of Nova Gorica in 2012. He is a research fellow at Faculty of Electrical Engineering (University of Ljubljana) and a member of User-adapted communications and ambient intelligence lab, as well as a research fellow at Scientific Research Centre of Slovenian Academy of Sciences and Arts. His research interests include affective computing, affect and emotions in human perception and cognition, cognitive semantics, human-computer interaction, and social intelligence.

Andrej Košir PhD, b.s.c math, received the Ph. D. degree in electrical engineering from the University of Ljubljana in 1999. Since 2014 he is a full professor at the Faculty of Electrical Engineering, University of Ljubljana, and a head of User-adapted communications and ambient intelligence lab. He is active in a broad research fields including: user modeling and personalization (user models, recommender systems), user interfaces (machine learning method design, statistical analysis), and social signal processing.