IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 October 2024, accepted 7 January 2025, date of publication 30 January 2025, date of current version 7 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3537156

RESEARCH ARTICLE

Automatic Classification of Difficulty of Texts From Eye Gaze and Physiological Measures of L2 English Speakers

JAVIER MELO[1], LEIGH FERNANDEZ[2], AND SHOYA ISHIMARU[3], (Member, IEEE)
[1]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[2]Psycholinguistics and Language Development, University of Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[3]Graduate School of Informatics, Osaka Metropolitan University, Sakai 599-8531, Japan

Corresponding author: Javier Melo (javier.melo@dfki.de)

This work was supported in part by Japan Society for the Promotion of Science under Grant 24K02962.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Team of DFKI.

ABSTRACT Reading is an essential method for adults to learn new languages, but difficulty reading texts in a foreign language can increase learners' anxiety. Identifying text difficulty from the reader's perspective can aid language learning by tailoring texts to readers' needs. There is little research focusing on L2 speakers or using a multimodal approach, i.e., using multiple sensors, to detect subjective difficulty. In this study ($N = 30$) we determined L2 speakers' subjective difficulty while reading using language proficiency and objective text difficulty, combined with physiological data. We compared machine learning classifiers combining eye, skin and heart sensor data against models using each modality separately. Additionally, we assessed the effect on model performance of shifting the data to account for delayed physiological responses. The models detected 3 levels of subjective difficulty (low, medium, high) and were evaluated using leave-one-participant-out (LoPo) and leave-one-document-out (LoDo) cross-validation. The results showed acceptable levels of generalization to new participants ($Acc_{LoPo} = 0.434$) and documents ($Acc_{LoDo} = 0.521$). Combining sensor data from all modalities improved predictions in both LoDo and LoPo cross-validation, compared to each modality in isolation. Shifting the data to account for physiological response delay did not improve model performance compared to not shifting the data. These findings support refining subjective difficulty detection models and their implementation in adaptive language learning systems. Finally, this work contributes to the field of cognitive science and technology by laying the foundation for innovative approaches to cognitive state detection.

INDEX TERMS Cognitive load, electrodermal activity, eye-tracking, human–computer interaction, L2 English speakers.

I. INTRODUCTION

Reading is one of the most important ways adults learn new languages. It is a complex task that requires the ability to not only understand the written words, but also to make sense of how they relate to each other. Difficulty reading a text in a foreign language can increase learners' anxiety about reading in the foreign language [1] and even discourage them from continuing to learn it. Moreover, simplifying texts has been found to improve readers' attitudes and recall [2], [3]. In this context, identifying difficulty from the reader's point of view provides an important opportunity to facilitate reading while promoting language learning.

The literature on text difficulty and its measurement is extensive. Research has shown that subjective or perceived difficulty during reading can be estimated by analyzing physiological biomarkers associated with cognitive load, which are captured using sensors [4]. The majority of

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda.

studies collecting sensor data during reading have focused on detecting cognitive measures other than subjective difficulty such as comprehension [5], [6], interest [7], distraction [8] and engagement [9]. Despite the importance of using data from multiple types of physiological activity [10], the studies focusing on subjective difficulty have only used data from the eyes [4], [11]. Also, of all of these studies involving reading in a first language (L1), none of them focused on second language (L2) speakers, i.e., language learners, who are known to display different reading patterns than native speakers [5].

The significance of studying L2 speakers is magnified by the historical emphasis on monolingual individuals in scientific research. In a world where bilinguals outnumber monolinguals, the focus on L2 speakers becomes critical. Their reading behaviors, shaped by interactions between languages, pose a number of challenges that differ from those of research focused on L1 speakers. For example, the study of L2 speakers may require consideration of factors such as language proficiency level in order to gain a more nuanced understanding of subjective text difficulty in the context of language learning.

In this work we extend previous research by measuring L2 speakers' subjective text difficulty using a multimodal approach that combines data recordings related to eye, skin, and heart activity. We recruited participants and asked them to read several paragraphs of text while we recorded physiological data. After each text, they were asked to rate the difficulty and to answer a comprehension question. They also took a language proficiency test. The features extracted from the sensor data were combined with the participant and text information to predict self-reported subjective difficulty, which served as the ground truth or label.

Of the reviewed studies that measured cognitive states, it is noteworthy that many did not use high-performance or medical-grade devices for their experiments. Instead, they chose more affordable devices that are closer to consumer-grade options (e.g., [7], [8], [9]). What these studies have in common is a practical approach, as opposed to a pure research approach. For example, the aim of Pai et al. [9] was to measure reader engagement in real time, which could be used to support students in a learning environment. Although they used the Tobii 4c, a low cost eye tracker, it could be argued that a high performance research system such as that used by Hollenstein et al. [6], as well as seating the participants at a fixed distance from the screen, might have resulted in better performing models. However, these methodological choices would limit the impact of such a system due to the increased price and reduced flexibility. Our focus was on language learners who may eventually use consumer devices such as smartwatches and gaming eye trackers to measure biomarkers. Therefore, our experimental design aimed to create an experience similar to that of using consumer devices, prioritizing external validity over the highest possible data quality.

The contributions of this paper are threefold. First, it contributes to the literature on subjective difficulty and cognitive load by investigating the effectiveness of a multimodal approach that integrates data from multiple sensors to provide a more comprehensive understanding of readers' cognitive and physiological experiences during text processing. Second, it extends the language learning literature by focusing on L2 speakers, a distinct group whose reading patterns and challenges differ from those of L1 speakers on average and thus require special attention. By taking language proficiency into account, this research advances our understanding of how learners' language skills affect their subjective ratings of text difficulty. Finally, our work contributes to the development of adaptable reading materials based on the learners' individual characteristics and needs, providing valuable insights for creating personalized and effective language learning resources.

II. RELATED WORK

A. TEXT DIFFICULTY

Tamor and colleagues distinguish between three types of text difficulty: text-related or objective, performance-related or behavioral, and reader-related or subjective [12]. Each type of difficulty is a slightly different indicator, so the choice to focus on one or the other depends on the objectives. In education, the ability to measure text difficulty is critical for selecting appropriate materials for students; finding materials that are hard enough to challenge the students but also easy enough to keep them motivated to keep going.

Objective difficulty operationalizes text difficulty in terms of characteristics such as word and sentence length, word frequency, and sentence cohesion, which involves the relationship between the elements of a text that affect its comprehension [13]. It is typically measured using a readability formula, such as the Flesch Reading Ease [14]. This formula takes into account the number of words, sentences and syllables in a text to generate a scale ranging from 0 to 100, with 100 being the highest readability score (i.e., the least difficult text). This and similar formulas assume that longer sentences and words are harder to read and understand. Another way to calculate objective difficulty is to analyze the predictability of words, understood as the likelihood of a word occurring given its context in a sentence. More predictable words should be easier to understand. A third way to measure objective difficulty is to calculate how common the words are in a particular text, relative to a large corpus, such as the SUBTLEX, a corpus of English media subtitles from the US and the UK [15], [16]. The underlying assumption is that common words are easier to understand than uncommon words.

Behavioral difficulty is determined by the reader's performance on a text [12]. For example, a difficult text elicits more errors than an easy text when read aloud [12], [17]. The behavioral difficulty of a text could also be calculated based on the performance of readers on comprehension questions.

In this case, it is important to consider the quality of this metric depends heavily on the quality of the questions. In addition to characteristics like age, native language, etc., behavioral difficulty can be used to determine the language level of the reader. Appropriate learning materials could be selected based on this level.

Objective and behavioral measures of text difficulty are widely used to select appropriate reading materials for learners, incorporating text characteristics and learner's performance. However, these types of measures have some limitations. One limitation is that they ignore the context of the reading situation. For example, a text of a given objective difficulty level might be perceived as easier if it is accompanied by a picture that's relevant to the text. Another limitation is that these measures ignore individual differences among readers, such as prior knowledge of the topic, language proficiency, current mood, interest, and motivations. For example, the same text about machines may be easier to read for a native speaker who is an engineer enjoying their vacation than for an L2 speaker who is a journalist having a bad day. Using subjective measures of reading difficulty can overcome some of these limitations.

Subjective difficulty operationalizes text difficulty in terms of what a given reader would consider ''easy'' or ''hard'' [12]. Typically measured by self-report, this judgment is influenced not only by objective text features perceived by the reader, but also by the reader's individual characteristics, such as reading ability, language proficiency, cognitive state, motivation, developmental level, prior knowledge, and the context of the reading. The cognitive effort that a reader puts into reading a passage is reflected in their rating of the text's difficulty. A valuable measure that reflects the use of mental resources and effort is cognitive load [10], [18], making it an excellent proxy for subjective difficulty.

Cognitive load, according to Paas et al. [19], ''represents the load that performing a particular task imposes on the cognitive system'' (p. 420). When faced with complex tasks, such as reading a difficult text, individuals tend to exhibit higher levels of cognitive load, and conversely, simpler tasks elicit lower levels of cognitive load [18]. Cognitive load has been measured using asynchronous methods, such as post-task questionnaires (e.g., [20]) and synchronous or real-time methods, such as sensors (e.g., [21], [22]) that provide physiological measures during the task.

B. PHYSIOLOGICAL SENSING OF TEXT DIFFICULTY

Subjective text difficulty can be indirectly measured using physiological biomarkers recorded by sensors. Ayres et al. [10] found that tasks of higher complexity, such as reading a difficult text, produce higher cognitive load than tasks of lower complexity. Furthermore, changes in cognitive load are reflected in the body and are therefore detectable by physiological measures [10]. These biomarkers could be analyzed to determine the subjective difficulty of the text that triggered the physiological response.

These biomarkers are collected from, among others, the eyes [4], the heart [23], the skin (e.g., electrodermal activity [7], skin temperature [24]) and the brain (e.g., electroencephalography [6]). Not all biomarkers are equally sensitive to changes in cognitive load. A review of the validity of physiological measures found that biomarkers related to the eyes were the most sensitive, followed by measures related to the heart and the lungs (i.e., the cardiovascular and respiratory systems, respectively), the skin, and the brain activity [10]. We will briefly describe the measures relevant to this work.

The main features that can be extracted from eye activity are those derived from pupil dilation, blinking and fixations [10]. Both pupil dilation and blink rate tend to increase as a result of increased cognitive demands. This increase occurs unconsciously and it is modulated by the action of neurotransmitters: norepinephrine in the case of pupil dilation and dopamine in the case of blink rate [25]. Fixations are usually defined together with saccades. Gaze points focused on a particular position are called fixations, and saccades are the jumps between fixations [26]. Unlike pupil dilation and blinks, fixations are more consciously modulated, but they have also been used to assess cognitive load [10]. Using gaze data, Reich et al. [4] were able to detect between high and low subjective difficulty. Their achieved performance leaving one participant out, that is, testing on a participant whose data was not used to train the model, was acceptable ($AUC = 0.71$).[1] On the other hand, their achieved performance when leaving one document out, that is, testing on a document unseen to the model, was slightly better than chance level ($AUC = 0.55$), indicating that the model uses document-specific information when determining its subjective difficulty.

Cardiac or heart-related activity can be measured to infer internal states, such as emotions [7]. Measures such as blood volume pulse (BVP), interbeat interval (IBI), heart rate (HR) and heart rate variability (HRV) can be obtained using plethysmography (PPG), a noninvasive technique [7], [27]. A recent study found high correlations between cognitive load and frequency and time domain measures of cardiac activity [27]. Frequency domain measures measure the power of heart rate variability in different frequency bands. For example, the low frequency (LF) band is measured between 0.04-0.15 Hz and has been found to be strongly correlated with cognitive load ($r = 0.91$). Time domain measures include the standard deviation of the IBI (SDNN) and the root mean square of the successive differences of the heartbeat intervals (RMSSD), which also showed high correlations with cognitive load (both $r = 0.7$), and the mean HR [27]. However, the usefulness of an indicator depends on the length of a task, as bias may occur if the task is shorter than the time required to compute the indicator reliably. For example, the minimum recommended period for measuring

[1]The AUC is the area under the ROC curve. A high AUC means that a model achieves a high true positive rate while maintaining a low false positive rate.

frequency-based HRV is 1 minute [28], whereas for time-domain measures, RMSSD measured at 10-second intervals has been found to correlate strongly with RMSSD calculated at 5-minute intervals ($r = 0.91$) [29]. For SDNN, the correlation to itself between 10 seconds and 5 minutes is only moderate ($r = 0.68$) [29].

Skin-related measures have also been found to be relevant to the detection of cognitive stress and arousal [10] and, to a lesser extent, to performance [24]. Changes in electrodermal activity (EDA) and skin temperature are modulated by the autonomic nervous system (ANS) [10], [30]. The ANS is responsible for defending the body against threats by increasing blood flow to the muscles, heart rate, sweat gland activation and respiratory rate, among other things [31]. In a fight-or-flight situation, skin conductance increases due to sweating [30] and skin temperature rises accompanied by the other bodily reactions [31]. The EDA signal is measured in micro-Siemens (μS) and it can be decomposed into tonic and phasic components [30]. The tonic component, also called the skin conductance level (SCL), represents the slowly changing skin conductance level. The phasic component contains information about the fast phasic pulses caused by the skin conductance response (SCR).

As we described so far, the ANS coordinates different organs to respond to threats. However, responses to stimuli take time to become observable in the body. For example, the heart needs time to make blood flow increase in the muscles. This would cause physiological measures to be delayed. To compensate for this, the data could be shifted to compensate for this delay. In this way, changes in sensor data would be observed as if they occurred simultaneously with the stimuli that elicited those responses. To properly shift the data, the response delays associated with each sensor must be known. This is the case for all the sensors described so far. A significant change in pupil diameter is typically observed 1.5 seconds after a word is processed by a reader [32]. For the heart activity, the response to a stimulus was found to be significant after 2 seconds [33]. For skin temperature, the delay was found to be about 4-5 seconds [34]. Finally, for EDA, the peak is reached between 1.5 and 6.5 seconds after a stimulus [35]. In this paper we explored the effects of shifting the data to compensate for the physiological delay in response.

While the problem of determining the difficulty of a text using sensors is challenging on its own, building this capacity into an application is even more so. Real-time difficulty detection provides a key opportunity to develop more useful learning materials. By incorporating this capability into a learning system, students can benefit from personalized and adaptive educational experiences.

C. COMPUTER-ASSISTED LEARNING

Computer-Assisted Learning (CAL [36]) systems have been developed for years under various names, including Computer-Aided Instruction [37], Intelligent Tutoring Systems [37] and Technology Enhanced Learning [38]. There are acronyms specific to language learning such as CALL and TELL, which add the word ''learning'' to the existing acronyms [39]. Throughout this work we refer to any computer-based system that assists or guides students in the learning process simply as a CAL. Most CALs include as part of their functionality some form of content adaptation based on a model of the learner. This model includes the learner's knowledge, which is then used to decide what to teach and how to teach it [36], [37], [40], [41]. By 2010, the most advanced reading-oriented CALs enhanced learning through interactive user experiences [42]. For example, they included hypertext, which links texts together and within themselves to support comprehension; hypermedia, which is visual and auditory information that complements text; and comprehension questions that test the user's knowledge. More recently, CALs have begun to consider the user's context in deciding what to teach [43]. Context includes location, time, the user's activity, and even cognitive states (e.g., emotions and motivation), all of which can be detected by using sensors [44].

Di Mitri and colleagues proposed a model to describe the process by which a CAL can integrate and interpret sensor data to provide real-time feedback to learners. They called it the Multimodal Learning Analytics Model (MLeAM, [44]). It consists of a cycle made of processes and outcomes. The MLeAM starts with multimodal data collection using sensors, annotating this data by defining ground truths or labels, training machine learning models to predict these labels, and providing timely feedback to the learner, who can change their behavior. This behavior is captured again with sensors, continuing the cycle.

The HyperMind is an example of a CAL that fits the MLeAM. It is an intelligent textbook that can be used to create learning materials that change in response to the reader's behavior as detected by sensors [26], [45]. For example, it could be customized to display an article that can be modified if the system detects a high level of subjective difficulty. In this way, if a reader is struggling with a particular paragraph of the document, HyperMind could respond by offering the learner a simplified version of the text to reduce its perceived difficulty [2]. The development of such adaptive systems is desirable and important, because it allows more students to learn at their own pace, with learning materials that match their abilities and needs.

D. THE PRESENT STUDY

In summary, there has been a lot of progress in the area of computer-assisted learning and, more recently, in the area of cognitive state detection during reading. Moreover, some studies have implemented multi-modal approaches, which we think are important to achieve better performance. However, little focus has been put on subjective difficulty in the context of language learning. From the studies reviewed, only one predicted subjective difficulty, but the only sensor data

they used was from eye tracking [4]. Regarding language proficiency, an important element when focusing on L2 speakers, none of the reviewed studies included language proficiency in their predictive models.

In this study, we determine the subjective difficulty of texts using a multimodal approach that combines data recordings related to eye, skin, and heart activity, in addition to language proficiency and objective text difficulty, which are particularly relevant in the context of language learning. To this end, we designed an experiment in a setup that collects participant's information and reading data, from which we extracted features. These features were used to predict subjective text difficulty. Both the detailed description of the experiment and the extracted features are presented in the following section.

We make the following hypotheses regarding prediction of subjective difficulty:

1) A multimodal approach, i.e., using a combination of several sensor data, achieves a higher performance compared to using data related to specific body parts in isolation.
2) Shifting the data to compensate for delays in physiological responses improves the model's performance, compared to using raw data.

III. METHOD

A. PARTICIPANTS

We recruited a total of 32 L2 speakers of English in Kaiserslautern, Germany, from the University of Kaiserslautern-Landau and the German Research Center for Artificial Intelligence. During preprocessing of the data, several trials were excluded based on quality criteria. For example, the data from 2 participants were not recorded properly and had to be excluded from the analysis. Detailed information about these exclusion criteria can be found in Section III-E1 dedicated to data preprocessing. The final sample consisted of 30 participants (15 male, 15 female). Their ages ranged from 21 to 32 years ($M = 25.3$, $SD = 2.49$). All reported normal or corrected-to-normal vision and were non-native English speakers, with age of acquisition of English ranging from 2 to 11 years old ($M = 5.89$, $SD = 2.77$). The most common languages reported to be spoken at home was Hindi ($N = 10$), followed by English ($N = 10$), Kannada and Spanish (both $N = 3$). The apparent contradiction that arises from reporting the age of English acquisition as 2 years or older while also reporting English as one of the languages spoken at home, could be explained by the fact that these participants either did not speak English at home until they began learning it in school or kindergarten, or that their exposure to the English language was not sufficient for them to consider themselves to have learned it (e.g., lexical code mixing [46]). See Table 1 for a summary of participant demographics and information about the languages they speak, which includes the results of an English proficiency test. For more details about this test, see the section dedicated to it 1.

TABLE 1. Demographic and language characteristics of the participants.

Variable	N	M	SD	Range
Age	30	25.3	2.49	21–30
Languages spoken	30	3.33	1.09	2–6
English learning age	29	5.89	2.76	2–11
English proficiency	30	77.67	11.01	55.75–93.75

B. APPARATUS

Eye movements were recorded using a Tobii Pro Fusion eye tracker sampling at 120 Hz and physiological measures were recorded using a wearable wristband, the Empatica E4, which has multiple sensors. Among them, an electrodermal activity sensor sampling at 4 Hz, an infrared temperature sensor sampling at 4 Hz and a photoplethysmography sensor sampling at 64 Hz [47]. Stimuli were presented on a Samsung SyncMaster 2443 24" flat panel LCD display (1920 × 1200 resolution at 60 Hz refresh rate). Participants were seated at a distance of 50-80 cm from the monitor, in a room illuminated by artificial light coming from above. The interaction with the participants was mediated by a website developed in Next.js [48]. This platform was used to present the stimuli and collect user data through clicks and keystrokes.

C. MATERIALS

1) READING STIMULI

The main task of the experiment consisted of one practice trial and 30 critical items presented in English. The critical items were selected paragraphs from the Provo corpus, which is a collection of 55 paragraphs that were annotated with predictability measures for each word in each sentence [49]. Each paragraph had two predictability measures associated with them. The first measure, "LSA context score", considers the predictability of words in the context of the sentence in which they appear, and the second measure, "LSA response match score", was obtained by analyzing data from an experiment in which participants were asked to predict the next word of each word in the sentences.

To ensure that the items covered a wide range of difficulty, we measured the objective difficulty of the 55 paragraphs from two sources: predictability and lexical difficulty. Predictability was calculated by taking the average of the z-scores of the two predictability measures provided by Luke et al. [49]. For lexical difficulty, we calculated the average word frequency in each text using the SUBTLEX corpus as a reference [15], [16], and then we calculated the z-score of it. To determine word frequency, we used "wordfreq" [50], which assigns a Zipf value to every word. The Zipf scale ranges from 1 to 7, where a word with the value of 1 is a very uncommon word, whereas a word with a Zipf value of 7 is a very common one. Finally, we calculated

TABLE 2. Descriptive statistics for objective difficulty indicators of the critical items.

			Text difficulty		
		Overall	Low	Medium	High
LSA context score	Mean	0.53	0.58	0.54	0.46
	SD	0.06	0.04	0.04	0.04
	Min	0.4	0.52	0.48	0.4
	Max	0.65	0.65	0.59	0.52
LSA response match score	Mean	0.24	0.26	0.24	0.22
	SD	0.05	0.03	0.06	0.05
	Min	0.14	0.21	0.14	0.15
	Max	0.34	0.3	0.34	0.33
Zipf frequency	Mean	5.17	5.31	5.21	4.99
	SD	0.26	0.23	0.26	0.2
	Min	4.69	4.94	4.87	4.69
	Max	5.61	5.61	5.59	5.34
Text length in words	Mean	50.63	52.3	50.2	49.4
	SD	4.84	4.52	5.45	4.5
	Min	39	46	39	42
	Max	59	59	58	55
Mean sentence length	Mean	21.17	20.53	20.03	22.93
	SD	4.61	5.58	4.49	3.43
	Min	12.5	12.5	13	17
	Max	29.5	29.5	26.5	27.5

the average between predictability and lexical difficulty to obtain the objective difficulty of each paragraph. We then sorted the list of paragraphs by objective difficulty and filtered this list to provide paragraphs with similar total length as well as sentence length, according to how readability formulas such as the Flesch Reading Ease determine text difficulty [14]. This way, the selected paragraphs would be similar in this regard, but different in terms of lexical difficulty and predictability. We then divided the list into three difficulty categories: low, medium and high, and selected 10 texts from each category, plus one for the practice trial. We created one comprehension question for each paragraph. Both the comprehension questions and the selected texts were reviewed by a panel of experts to ensure that the questions and the texts were clear and that the assigned difficulty levels were appropriate for the selected texts. See Table 2 for information about the items and see Appendix for the list of the critical items used in this study.

2) LexTALE

To measure the participants' English proficiency, we used the LexTALE test, which was designed for L2 speakers [51]. The test asks the participants to distinguish between words and pseudowords in order to calculate a score based on their lexical knowledge, i.e., how many words they know. This score can be interpreted to distinguish between lower intermediate, upper intermediate and advanced users. The score distribution of the participants is shown in Figure 1. The red vertical lines indicate how to interpret the scores as CEF English proficiency levels [51]. Scores below 60 correspond to the B1 level, which is a lower intermediate level or below. Scores between 60 and 79 correspond to B2, which is an upper intermediate level. Scores of 80 and above are considered C1 and C2, which includes lower and upper advanced/proficient users.

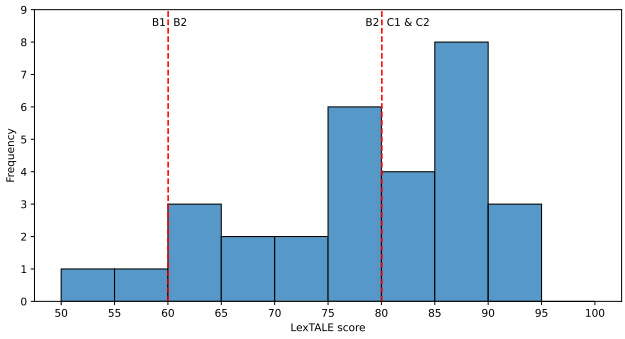


FIGURE 1. Histogram of the LexTALE scores with red vertical lines indicating the levels of English proficiency [51].

3) LANGUAGE AND SOCIAL BACKGROUND QUESTIONNAIRE

We asked the participants to complete an adapted version of the Language and Social Background Questionnaire [52] to collect demographic information about the participants and their language background. Language background was particularly important to ensure that the participants were L2 speakers and not native English speakers.

D. PROCEDURE

The experiment consisted of five parts. First, participants were informed about the task and were asked to read and sign a consent form. Second, participants completed the LexTALE [51] to test their English proficiency. Third, they performed the reading task, which was the main task of the experiment. During this part, sensor data was recorded. Fourth, they answered the questionnaire on language and social background [52]. Finally, as a reward for their participation, they were asked to choose between receiving 10 € in Amazon credit or one participation hour.

The reading task consisted of 1 practice trial followed by 30 critical items. Before starting, the participants put on the Empatica E4 wristband and performed a 9-point calibration of the Tobii Pro Fusion eye tracker. Every 10 trials, the participants were asked to take a break and to recalibrate the eye tracker to ensure the best possible accuracy. They were allowed to take breaks between trials as well, but they had to recalibrate before continuing with the task.

Each trial consisted of 5 parts. First, they were instructed to look at a fixation cross located in the upper left part of where the first letter of the paragraph would be. Second, after 1.5 seconds, the fixation cross was replaced by a paragraph that they had to read for comprehension. They were asked to press the space bar when they finished reading the paragraph, so that they could move on to the next part of the task. Third, they had to rate the difficulty of the text on a scale of 1 to 5, with 1 being the easiest and 5 being the most difficult. To do this, they had to move a slider that started in

the middle by default, indicating medium difficulty. Fourth, to encourage them to read the texts carefully, they answered a comprehension question, which was a true/false statement about the text that could be answered by pressing "j" for true or "f" for false. Fifth, they were presented with the text again and were asked to click on the words they did not know. After pressing the space bar one last time, they either proceeded to the next trial or went to the pause screen to take a break and recalibrate. See the Appendix for the list of texts and their corresponding comprehension questions.

Except for the practice trial, the order of presentation of the texts was randomized for each participant. Each paragraph was presented as a whole in the center of the screen but flushed to the left in a monospaced font (Courier New), 24px in size, with a triple-spaced line height. This was done to make the paragraphs easy to read and to make it easy to assign fixations to each word. See Figure 2 for an example critical item.

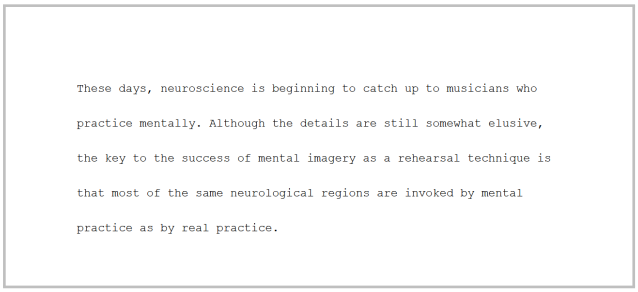


FIGURE 2. Visualization of one of the critical items.

E. ANALYSIS

All data analyses were performed using Python, a general-purpose open-source programming language [53]. To extend Python's functionality, we used several packages. These include "pandas" [54], to work with datasets, "matplotlib" for visualization [55] and "scikit-learn" for machine learning [56].

1) PREPROCESSING

During preprocessing, invalid participant and trial data were removed from the dataset. First, gaze data was filtered to exclude trials based on these criteria: (1) there were not enough gaze points data to calculate fixations (132 trials) (2) the ratio of fixations to words was less than 0.5, indicating that less than a half of the words in the text were read (9 trials), (3) visual inspection of the gaze path indicated incomplete readings (2 trials). The data of the Empatica E4 also had 2 missing trials from temperature and EDA respectively. After preprocessing, the data set contained 817 valid trials from 30 participants (85% of the expected 960 trials).

For the analysis, data related to each participant, i.e., sensor data from the eye tracker and the wristband, trial information such as start and stop timestamps, identifier of the text being read, text difficulty ratings, responses to comprehension questions, and LexTALE responses, were combined to create data sets, each one with their own type of data (i.e., sensor data or trial information), timestamps, and participant and trial identifiers. To make the data sets compatible with each other, the time stamps were adjusted to be in the same time zone. Then we prepared each data set separately corresponding to each data source: participant input (i.e., comprehension questions, difficulty ratings and LexTALE), gaze data, cardiac data and skin data.

The comprehension questions were interpreted by comparing the response to the correct answer. The average accuracy rate per participant was 78% (standard deviation was 0.079), indicating that they read the texts carefully. The subjective difficulty rating of each text was measured on a 5-point scale with increasing difficulty. From left to right, the points were labeled as follows: (1) very easy, (2) easy, (3) medium, (4) difficult, (5) very difficult. This ordinal scale allows us to rank the texts in terms of their subjective difficulty, but the scale does not ensure that the intervals between the points are equal. For example, a participant might rate a large number of texts as "easy", but only rate the extremely easy texts as "very easy". We grouped the scale scores into three categories of subjective difficulty: "low", including the "easy" and "very easy" ratings, "medium", and "high", including the "difficult" and "very difficult" ratings. By grouping the subjective difficulty ratings in this way, it is possible to treat the task of predicting subjective difficulty as a classification problem. This allows us to visualize the correctly and incorrectly classified samples in a confusion matrix, and makes our results more comparable to the previous literature, which mostly treats the prediction of subjective measures as a classification problem as well. The LexTALE was scored by using the $\%\ correct_{av}$ (averaged % correct) described in Lemhöfer and Broersma [51]:

$$\frac{(N_{correct\ words}/40 * 100) + (N_{correct\ nonwords}/20 * 100)}{2}$$

The gaze data consisted of a data frame describing the participant's eye movements as detected by the eye tracker in the form of gaze points. Each gaze point had its associated timestamp and coordinates of where it was on the screen. Although most of the gaze points were located on the text, some points were on other parts of the screen. To focus on only reading, we kept only the gaze points located within the region where each text was presented in each trial. Then, we detected fixations and saccades from the gaze points using the approach proposed by Buscher et al. [57] and stored them in a data frame where each fixation and saccade was associated with its own time stamp. A screenshot of one of the trials is shown in Figure 3. The dots in the image represent fixations, which become larger as their duration increases, and the lines connecting the dots represent the saccades. The first fixation on the text is marked in green and the last fixation is marked in red. After detecting the fixations and saccades of each trial, we matched each fixation to the closest word in the text being read, taking the center of the word as

its position. We assumed that the closest word to a fixation was the word being read. This is important because some of the features rely on the specific words being read.

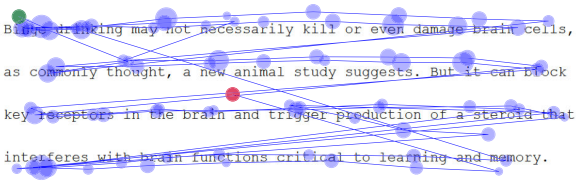


FIGURE 3. Visualization of fixations and saccades of a participant reading on one of the paragraphs.

To the data frame of fixations and saccades, we added the pupil diameter to each fixation. This was done by calculating the median of all the pupil sizes associated with the gaze points that make up these fixations. We calculated a baseline to perform a baseline correction, which is useful for examining changes in pupil size by reducing the impact of unwanted fluctuations in pupil size [58]. We did this by subtracting a baseline value from the pupil size values. This baseline value came from the median pupil size of the 1.5 seconds before each trial, which is the time when a fixation cross is being displayed. Considering that the processing of a word takes about 1.5 seconds to be reflected in a change in pupil diameter [32], we calculated shifted pupil diameters which are shifted in 1.5 seconds. This way, any change in pupil diameter caused by a difficult word or phrase appears in the data as if it occurred immediately.

We used the "cvxEDA" Python package [59] to decompose the raw electrodermal activity (EDA) signal of the skin into its tonic and phasic components. Additionally, we calculated a sparse indicator of the sudomotor nerve activity driver of the phasic component, also known as the "phasic peaks" [7], also using the "cvxEDA" package (see Figure 4).

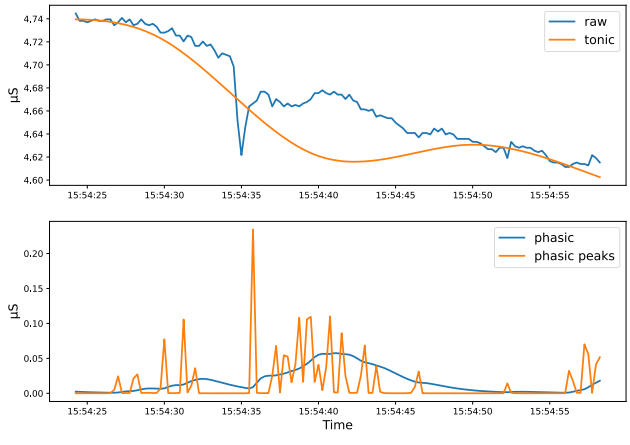


FIGURE 4. Visualization of skin conductance components over time for one of the trials. The top panel shows raw skin conductance and the tonic component. The bottom panel shows the phasic component and the phasic peaks.

To illustrate the effect of data shifting, we manually selected 10 seconds of a trial to visualize pupil size and the phasic component of EDA before and after the shift, which accounts for the delayed physiological responses. Figure 5 shows how pupil diameter and the phasic component of EDA align after the shift. Since both measures detect cognitive load based on different biomarkers, this effect is not clearly visible in most trials. However, in this example it is visible, which helps to show what this shift is intended to do.

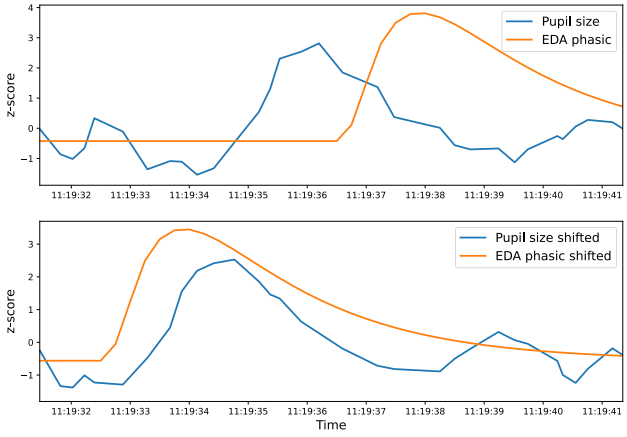


FIGURE 5. Visualization of pupil size and EDA's phasic component before delayed response correction (above) and after (below). Interval manually selected to illustrate the effect.

F. FEATURE ENGINEERING

We computed several features using all the data gathered during the experiment, including the sensors and the surveys. We hypothesized that shifting the data to account for physiological delayed response would improve the model's performance, compared to using the dataset with no adjustment for delayed response. To test this hypothesis, we computed 2 sets of features: the first one used the raw data (without shifting), and the second used shifted data.

From the gaze data, we calculated features derived from pupil diameter, and from fixations and saccades. From pupil diameter we calculated the mean, standard deviation, slope of the best fitting linear regression, and mean and standard deviation of the rate of change of the pupil diameter. Additionally, we calculated the mean pupil diameter corrected by subtracting the baseline. To account for delayed physiological response of the pupil, we calculated the same features again but shifting pupil size data by 1.5 seconds.

From the fixations and saccades, we calculated the mean, maximum, and standard deviations of fixation duration, saccade length, saccade speed and saccade angle. Saccade angle is determined by the line connecting the previous fixation to the current one. These angles are measured using a unit circle, where the starting fixation is at the center and the current fixation is on the circumference. The angle values begin at 0 on the right side of the unit circle, increasing in an anti-clockwise direction and decreasing in a clockwise

direction, culminating at plus or minus π radians (equivalent to 180°) on the left side. Additionally, as a measure of work skipping, we included word omission rate as defined by [6] which is calculated by counting the number of words that were not read (i.e., with no fixations assigned to them) divided by the total number of words. Finally, we added the following features used by Garain et al. [11]:

- Number of forward gaze points (Pf): number of words that were read for the first time (forward reading).
- Number of backward gaze points (Pb): number of words read again (backward reading)
- Duration of forward reading (Df): total duration of forward reading fixations.
- Duration of backward reading (Db): total duration of backward reading fixations.

From cardiac data, we used the "neurokit2" package [60] to calculate several indicators, starting with mean heart rate. From inter-beat intervals (IBI) we calculated the mean, median, standard deviation (SDNN) and root mean squared successive differences between them (RMSSD) [27]. To measure heart rate variability we calculated long term heart-rate variability (SD2), and the mean and standard deviations of the first and second grade derivatives (rate of change) of inter-beat intervals [61]. To consider delayed physiological response, we shifted cardiac data by 2 seconds [33] before calculating the features for the shifted dataset.

From skin temperature, we calculated the mean, standard deviation, slope, and the mean and standard deviation of its rate of change. For the baseline-corrected features, we used the temperature values divided by the baseline. For the shifted features, we shifted temperature data by 4.5 seconds to account for delayed physiological response [34].

Regarding electrodermal activity (EDA) we calculated the following features: the first and second derivatives to measure rate of change of both the tonic and phasic components of EDA. From the phasic component we calcualted the minimum, maximum and mean amplitude (in [7]), the sum of all positive EDA changes [35], and the number of phasic responses, operationalized as the number of phasic peaks observed, which we calculated by using the SciPy package for Python [62]. For our shifted features set, we shifted EDA values by 4 seconds, which is in the middle of the range reported by Leiner et al. [35].

In addition to all these features, we added the LexTALE score (language proficiency) and the objective difficulty of the text to the list of features. The full list of features can be found in Table 3.

G. EVALUATION PROTOCOL

To evaluate our approach, we trained machine learning models using Support Vector Machines (SVM) and Random Forest (RF) classifiers to predict subjective difficulty based on the features extracted from the trials. The feature extraction was done using a sliding window approach [63]. From each trial of each participant, we extracted the features from Table 3 using a 10-second sliding window, moving 1 second per iteration. The features calculated in each trial were assigned to one of the three classes (i.e. low, medium, high) corresponding to the subjective difficulty level of the text in the trial.

TABLE 3. List of all features.

Source	No.	Feature
Gaze	1–3	{mean, max, SD} of fixation duration
	4–6	{mean, max, SD} of saccade length
	7–9	{mean, max, SD} of saccade speed
	10–12	{mean, max, SD} of saccade angle
	13–14	Number of {forward, backward} fixations
	15–16	Duration of {forward, backward} reading
	17	Word omission rate
	18–20	{mean, SD, slope} of pupil diameter
	21	Mean pupil diameter corrected for baseline
	22–23	{mean, SD} of 1st derivative of pupil diameter
Heart	24	Mean heart rate
	25–28	{mean, median, SD, RMSSD} of IBI
	29	Long term heart-rate variability
	30–31	{mean, SD} of 1st derivative of IBI
	32–33	{mean, SD} of 2nd derivative of IBI
EDA	34–36	{min, max, mean} of phasic component of EDA
	37	Number of EDA phasic peaks
	38–41	{mean, SD} of 1st and 2nd derivatives of tonic EDA
	42–45	{mean, SD} of 1st and 2nd derivatives of phasic EDA
Temperature	46–48	{mean, SD, slope} of skin temperature
	49	Mean skin temperature corrected for baseline
	50–51	{mean, SD} of 1st derivative of skin temperature
Participant	52	LexTALE score
Text	53	Objective difficulty of the text

We used a leave-one-out cross-validation approach to separate training and testing data for classification. This method evaluates how well a model can predict labels from categories that were not included in training. For example, in the leave-one-document-out (LoDo) cross-validation, the model is first trained on all documents, except one, which serves as a test case. This process is then repeated, each time using a different document for testing, while the remaining data is used for training. A similar approach, known as leave-one-participant-out (LoPo), iteratively leaves out one participant for testing while using the others for training. This process continues until all documents or participants have been cycled through or until a specified number of iterations have been completed. The overall performance of the model is then evaluated based on the labels predicted in each iteration.

To test our hypotheses, we trained several machine learning models and evaluated them using LoDo cross-validation, where each paragraph was a document, and LoPo cross-validation. According to our first hypothesis, combining

features from multiple sensors should yield better results than training models using only features related to one part of the body in isolation. To test this hypothesis, we trained a full model for comparison, that included objective measures of text difficulty and English proficiency, along with the data from all the sensory modalities we collected. We then created variations of the full model that differed only in that each variation retained features from only one modality. This resulted in five conditions: (1) all physiological data, (2) gaze data only, (3) heart data only, (4) electrodermal activity data only and (5) temperature data only. According to our second hypothesis, there should be a difference between models trained on raw data, compared with the models trained on data shifted to account for delayed physiological response. Therefore, we had 5 × 2 conditions to test with LoDo and LoPo cross-validation, using SVM and RF to train machine learning classifiers. That was a total of 40 models.

We used accuracy (correct predictions divided by all predictions) and the F1 score to measure the performance of the models. The F1 score is the harmonic mean of the precision and recall, and it is defined by the expression:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Since the data is not balanced as seen in Figure 6, we used the F1 with weighted average, provided by the "Scikit-learn" Python package [56], which is suitable for unbalanced multi-class classification.

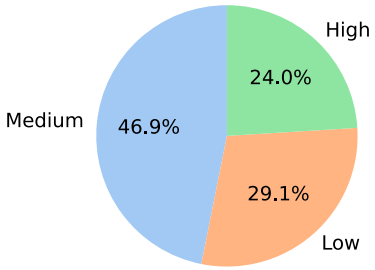


FIGURE 6. Class distribution of subjective difficulty.

In addition to the model performance, we provided a confusion matrix for the overall best model in the LoPo cross-validation, and the best model in the LoDo cross-validation condition. A confusion matrix is a visual representation of the predictions made by a classification model. It displays correctly and incorrectly classified data in a matrix form [64]. Figure 7 shows the confusion matrix for a 3-class subjective difficulty classification problem. The cells following the green diagonal line indicate correct class predictions. For example, TP_M represents texts of medium subjective difficulty that were correctly classified as having a medium difficulty by the model. All other cells represent incorrect predictions. For example, texts of low subjective difficulty that were incorrectly classified as having a high difficulty are represented by E_{LH}.

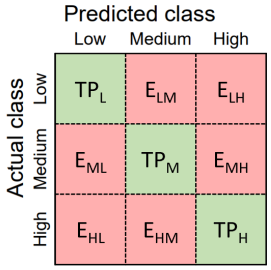


FIGURE 7. Illustration of confusion matrix for a 3-class subjective difficulty classification problem.

TABLE 4. Model performance using LoPo cross-validation.

	SVM		RF	
	F1	Accuracy	F1	Accuracy
All modalities				
Raw	**0.434**	0.434	0.411	**0.451**
Shifted	0.433	0.433	0.400	0.435
Gaze only				
Raw	0.417	0.416	0.409	0.441
Shifted	0.411	0.409	0.400	0.429
Heart only				
Raw	0.383	0.385	0.424	0.436
Shifted	0.394	0.396	0.397	0.411
EDA only				
Raw	0.393	0.402	0.385	0.402
Shifted	0.393	0.402	0.406	0.417
Temperature only				
Raw	0.397	0.394	0.410	0.424
Shifted	0.380	0.378	0.403	0.422

TABLE 5. Model performance using LoDo cross-validation.

	SVM		RF	
	F1	Accuracy	F1	Accuracy
All modalities				
Raw	**0.521**	**0.521**	0.498	**0.521**
Shifted	0.514	0.515	0.492	0.515
Gaze only				
Raw	0.478	0.477	0.480	0.498
Shifted	0.487	0.487	0.486	0.502
Heart only				
Raw	0.420	0.418	0.503	0.514
Shifted	0.415	0.415	0.473	0.487
EDA only				
Raw	0.393	0.403	0.470	0.484
Shifted	0.387	0.397	0.467	0.480
Temperature only				
Raw	0.452	0.432	0.478	0.485
Shifted	0.429	0.432	0.459	0.469

IV. RESULTS

A. HYPOTHESIS TESTING

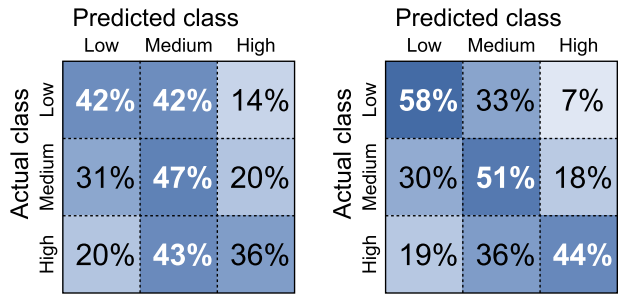
Table 4 show the performance of the models evaluated using LoPo cross-validation, whereas Table 5 uses all features. All models were evaluated using weighted F1, and the best performing model is highlighted in bold in each table.

Tables 4 and 5 show that the best performing models in the LoPo and LoDo conditions were the models that integrated data from all the sensors, supporting our Hypothesis 1, which is based on the Multimodal Learning Analytics Model (MLeAM, [44]).

Across Tables 4 and 5, the best performing models consistently emerged from the raw data. This suggests that shifting the sensor data to account for the response delay in each modality does not necessarily improve model performance. The only exceptions were in the LoDo condition using gaze features and on the SVM models in the LoPo condition using cardiac features. However, the increase was about 0.01 points in both F1 and accuracy. Considering these results, there is not enough evidence to support our Hypothesis 2, which expected models using shifted data to outperform models using raw data.

B. MODEL PERFORMANCE

Figure 8(a) shows the confusion matrix summarizing the predictions of the best performing model, the SVM classifier, in the LoPo cross-validation condition ($Acc = 0.432$). This was selected over the model using Random Forest, because of having a more balanced confusion matrix, explained by its higher F1 thttp://192.168.43.56:8501/score. The model was trained with SVM on the full raw dataset, which includes gaze, cardiac, electrodermal activity, and temperature features, in addition to the common features used in all other conditions, i.e. objective text difficulty and English proficiency score. On the LoDo condition, Figure 8b shows the confusion matrix summarizing the predictions of the best performing model ($Acc = 0.521$), which also used the full raw dataset. Because SVM and RF achieved the same accuracy, the SVM model was chosen, as it obtained the highest F1 ($F1 = 0.521$).



(a) Leave-one-participant-out CV. (b) Leave-one-document-out CV.

FIGURE 8. Confusion matrices of the best model for each evaluation.

In Figure 8(b), showing the best model using LoPo cross-validation, the model correctly classified the subjective difficulty of a text about a 40% of the time, with the best performance being 47% for correctly identifying texts of medium difficulty and the lowest being 36%, for correctly classifying high-difficulty texts. Most of the classification confusion was observed with low and high difficulty texts being classified as of medium difficulty by the model. On the other hand, the model only classified low-difficulty texts as having high difficulty a 14% of the time and high-difficulty texts were classified as having low difficulty 20% of the time. The performance increases in LoDo cross-validation, as shown in Figure 8b. The model correctly classified the subjective difficulty of a text between 44% (high difficulty) and 58% (low difficulty) of the time. As in LoPo, most of the classification confusion was observed with low and high difficulty texts being classified as of medium difficulty by the model. On the other hand, the model only classified low-difficulty texts as having high difficulty a 7% of the time and high-difficulty texts were classified as having low difficulty 19% of the time.

C. FEATURE ANALYSIS

To gain a better understanding of the performance in the LoPo and LoDo condition, we made a complementary analysis of the features, focusing on the participants and in the documents respectively. Figure 9 shows the Pearson's correlation of the features with the subjective difficulty for each participant and Figure 10 shows the Pearson's correlation of the features with the subjective difficulty for each text. Negative correlations are shown in shades of blue and positive correlations are shown in shades of red. Both ends are connected by 0, which is shown in gray.

All features in Figures 9 and 10 show some degree of either positive or negative correlation with subjective difficulty. However, the light colors indicate that most of the correlations are weak and could be explained by chance or by being participant- or document-specific. To identify the true correlations, or those that are independent of specific participants or documents, we looked at the homogeneity of these correlations across participants and documents.

In Figure 9, objective difficulty is positively correlated with subjective difficulty for almost all participants. This means that this feature is user independent, as it shows a positive correlation with subjective difficulty regardless of the participant. The same cannot be said for mean pupil diameter. This feature looks heterogeneous, with some correlations negative across participants, many around 0, and some positive. This suggests that the mean pupil diameter is user dependent, that is, its correlation with subjective difficulty varies from participant to participant.

Regarding the correlations between features and subjective difficulty separated by document, Figure 10 shows that the features are more heterogeneous than in Figure 9. For example, the standard deviation of the rate of change of pupil diameter (''pupil_diameter_d1_std'') is positively correlated with subjective difficulty in some documents but negatively correlated in others, which would make this measure more document dependent than the other features. However, positive EDA count seems to be more consistently positively correlated with subjective difficulty and omission rate seems to be more consistently negatively correlated with subjective difficulty, suggesting that this feature is

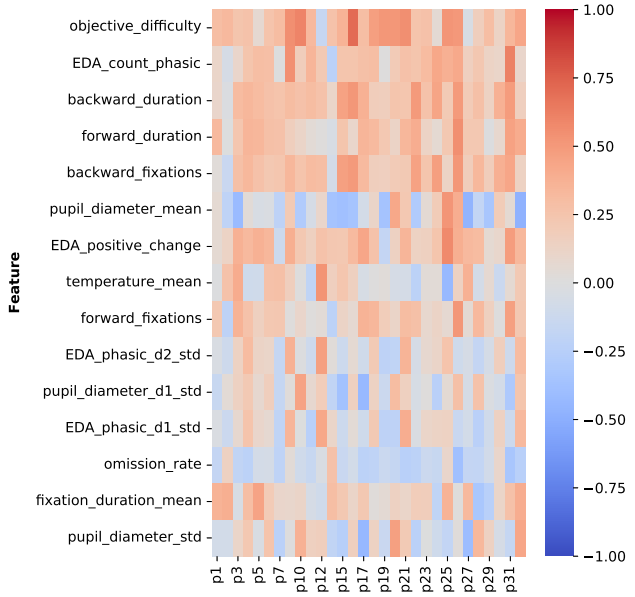


FIGURE 9. Pearson correlation between subjective difficulty and the 15 most correlated features separated by participant (raw data). Note: feature names with d1 and d2 refer to first and second derivatives (rate of change) respectively.

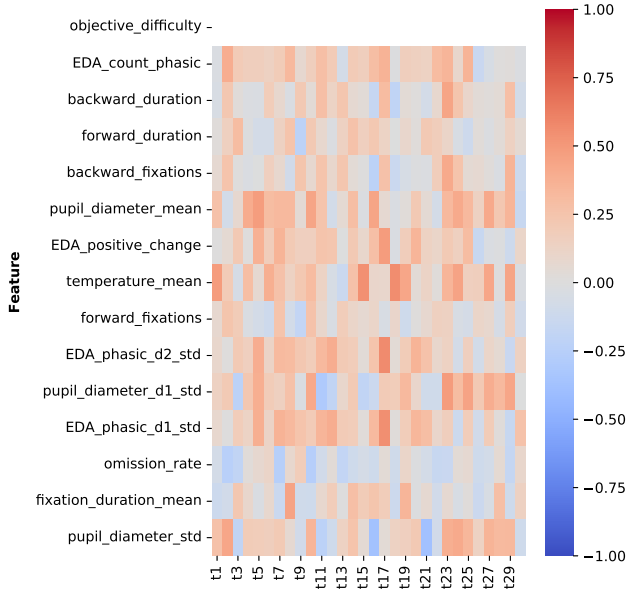


FIGURE 10. Pearson correlation between subjective difficulty and the 15 most correlated features separated by document (raw data).

more text independent than the other features. Note that the objective difficulty row is empty. This is because the correlation between objective difficulty and subjective difficulty is undefined since objective difficulty is constant within documents.

To complement the feature correlation analysis, we computed the Permutation Feature Importance (PFI, [65]) to determine the importance of the top 10 features in predicting

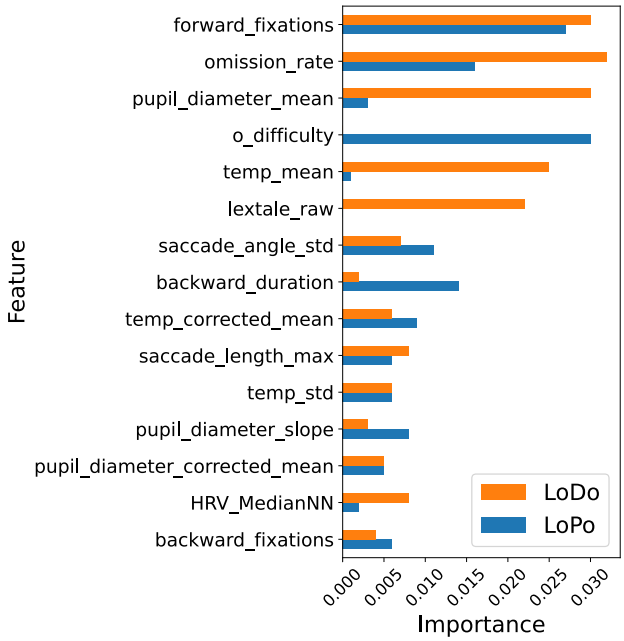


FIGURE 11. Permutation importance of the features in the LoDo and LoPo cross-validation conditions (only first 15).

subjective difficulty. This measure of feature importance indicates the decrease in the model performance when a feature is shuffled. The higher the importance of a feature, the worse the performance of the model after shuffling that feature. We computed the PFI separately for the LoPo and LoDo cross-validation conditions. For each step of each cross-validation process, a new model was trained and evaluated. If a model performed better than chance, we considered it useful enough to predict subjective difficulty. Thus, we computed the PFI for that model 10 times and the calculated the average importance per feature. Then, using all the models for which we calculated the PFI, we calculated the average of the importance of their features per cross-validation condition. This process yielded the final ranking of features and their corresponding importance in predicting subjective difficulty.

Figure 11 shows the permutation importance of the features in LoPo and LoDo conditions, limited to the 15 features with the highest importance combining both conditions. The figure shows that features such as objective difficulty have a high importance in LoPo, as expected, given its homogeneity across participants as seen in Figure 9. On the other hand, the mean pupil diameter achieved a low feature importance in LoPo compared to its feature importance in LoDo, which could be explained by its heterogeneity across participants, as seen in Figure 9.

The difference in importance between LoPo and LoDo of "LexTALE'_raw" and objective difficulty seems unusual. LexTALE has a high importance in LoDo, meaning that its removal would decrease model performance, but a zero importance in LoPo, meaning that it does not affect model

performance at all when that validation method is used. Conversely, objective difficulty has a high importance in LoPo but a zero importance in LoDo. This can be explained by the fact that LexTALE remains constant within each participant and objective difficulty remains constant within each text. Thus, it makes sense that shuffling LexTALE does not change model performance in LoPo and shuffling objective difficulty does not change model performance in LoDo.

V. DISCUSSION AND CONCLUSION

In this paper we detected the subjective difficulty of texts while reading, using a multimodal approach that combined data recordings related to eye, skin, and cardiac activity. We designed a reading experiment for which we recruited participants from whom we collected reading data. We extracted features from these data to train machine learning models capable of predicting subjective text difficulty.

Our results indicate that subjective text difficulty can be detected using our approach, with acceptable levels of generalization to new participants and better levels of generalization for new documents. We found that using a combination of sensors improved the prediction of subjective difficulty, compared to using a single modality for prediction. Another finding was related to the delay in physiological responses. Shifting the data to account for these delays was not found to improve model performance.

Our results agree partially with the studies related to the prediction of subjective difficulty, or more broadly, to cognitive load. Reich et al. [4] modeled the prediction of subjective difficulty as a 2-class problem (high versus low difficulty) and reported an acceptable AUC score of 0.71 for LoPo and a near-chance AUC of 0.55 for LoDo. Our results in both LoPo and LoDo were at a similar level we consider ''acceptable''; higher than chance level but not high enough to be rated as good. The data used by Reich et al. consisted in reading data of 95 participants who read four texts of 5–6 pages each, while our data consisted in only 30 participants reading 30 short paragraphs. It is unclear why Reich et al. achieved a lower performance on LoDo compared to LoPo, whereas our results show the opposite. One possible explanation for this difference is that we included participants' language proficiency in our models, which was one of the most important features for LoDo. Despite Reich et al. mentioning that some of their participants were L2 speakers of English, they did not include any measure of language proficiency in their models [4].

Our results are inconsistent with previous literature studying delayed physiological responses [32], [33], [34], [35]. These studies provide specific time delays associated with the stimulus responses of different body parts, which could be used to shift the sensor data, thereby improving models trained with these data. However, shifting data did not improve our models, suggesting that either the effect of this delay on performance is minimal or the specific characteristics of this study did not allow an effect to be observed. We analyzed the data using a sliding window of 10 seconds. It is possible that the effect of shifting the data would only be significant with smaller window sizes. On the other hand, following previous research on physiological delayed response for each modality [32], [33], [34], [35], the data were shifted by constant intervals to account for these delays. However, this ignores individual differences that may have influenced the results. For example, one participant's pupils may begin to dilate in response to stimuli earlier than another participant's pupils, so the data should have been shifted by a different interval for each participant. Techniques such as dynamic time warping may be useful for aligning data from multiple modalities. To this date, we have not found any other study that has shifted the data as suggested by these authors or used alignment techniques to improve the prediction of participants' cognitive states. Therefore, future studies should continue to explore this hypothesis in different scenarios.

As we already mentioned in the introduction, one of the goals of our research design was to create a system that would be realistic and affordable to implement. For this reason, we chose the Empatica E4 and the Tobii Pro Fusion. One of the disadvantages of this choice is that the data collected may not have been as good as if it had been collected with more accurate (albeit more expensive) equipment. We have provided many examples of studies that have also used low-cost devices and obtained good results. In addition, some validation studies have reported high correlations between data collected with the Empatica E4 and data from medical-grade devices ([66], [67], [68])) and one study achieved similar performance by comparing low-cost devices with a more accurate one [68]. However, high correlations with a gold standard do not always guarantee reliable predictions based on the data. In addition, we have not found any scientific articles analyzing the validity of the data recorded by the Tobii Pro Fusion. For these reasons, we recommend caution when interpreting the results obtained from these devices, but we also encourage more research, to evaluate the accuracy and predictive power of these devices in different contexts.

One implication of this study relates to the fact that our participants were L2 speakers of English. We found language proficiency to be an important feature for predicting subjective text difficulty in L2 speakers. However, it remains to be investigated whether the inclusion of a language proficiency indicator such as the LexTALE would be as relevant when the participants are L1 speakers. This is because it is known that L2 speakers do not read in the same way as L1 speakers and because the variance of L1 speakers' language proficiency scores may be lower than the variance of L2 speakers' language proficiency scores, reducing the ability of such a measure to contribute to the model performance.

Building on the insights gained in this study, we propose two directions for extending the literature on subjective text difficulty prediction. The first direction involves the

TABLE 6. List of items.

ID	Text	Comprehension Question	Correct	Difficulty
1	There are now rumblings that Apple might soon invade the smart watch space, though the company is maintaining its customary silence. The watch doesn't have a microphone or speaker, but you can use it to control the music on your phone. You can glance at the watch face to view the artist and title of a song.	Apple had already launched the Apple Watch.	False	Low
2	The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity. Heat reactions usually occur when large amounts of water and/or salt are lost through excessive sweating following strenuous exercise. When the body becomes overheated and cannot eliminate this excess heat, heat exhaustion and heat stroke are possible.	The human body can get overheated to a point where it can be life threatening.	True	Medium
3	It was a forbidding challenge, and it says much for Winstanley's persuasive abilities, not to mention his self-confidence, that the admiralty agreed to fund him. No lighthouse had ever been built out to sea on an isolated rock before. The Eddystone is hard enough to avoid, but landing on it is a separate challenge entirely.	Despite the challenge of building a lighthouse on a rock, Winstanley got funding from the admiralty to do it.	True	Medium
4	A clergyman remarked to him, "The Lord is on our side." "I am not at all concerned about that," replied Mr. Lincoln; "for I know that the Lord is always on the side of the right. But it is my constant anxiety and prayer that I and this nation should be on the Lord's side."	Mr. Lincoln believes that the Lord is on his side.	True	Low
5	Some months later, Michael Larson saw another opportunity to stack the odds in his favor with a dash of ingenuity. He walked into his bank one day and asked to withdraw his entire account balance, but with an unusual stipulation: He wanted as much of the cash as possible in one dollar notes.	Michael Larson's unusual stipulation was that he wanted as much of the cash to be handed out the same day.	False	Medium

TABLE 6. *(Continued.)* **List of items.**

6	Greg Anderson, considered a key witness by the prosecution, vowed he wouldn't testify when served a subpoena last week. His lawyers said he was prepared for a third prison stay to maintain his silence. Anderson was released July 20 after a two-week stay for previously declining to testify before a different grand jury.	Despite being a crucial witness in the case, Greg Anderson refused to testify.	True	High
7	Steam sterilization is limited in the types of medical waste it can treat, but is appropriate for laboratory cultures and substances contaminated with infectious organisms. The waste is subjected to steam in a sealed, pressurized chamber. The liquid that may form is drained off to the sewer or sent for processing.	Steam sterilization is appropriate for all kinds of medical waste.	False	High
8	Owls are more flexible than humans because a bird's head is only connected by one socket pivot. People have two, which limits our ability to twist, Forsman added. Owls also have multiple vertebrae, the small bones that make up the neck and spine, helping them achieve a wide range of motion.	Humans have as many neck sockets as owls do.	False	Medium
9	Even in the same animal, not all bites are the same. A new study finds that because the force in a muscle depends on how much it is stretched, an animal's bite force depends on the size of what it is biting. The finding has direct implications for ecology and evolution.	The bite of the animal depends on what is being bitten.	True	Low
10	Interestingly, the heaviest isotopes physicists have managed to synthesize so far don't behave quite like science's best current models predict, so stable superheavy nuclei are likely to be full of surprises. Chemists cannot even predict with any certainty whether these materials will exist as gases, liquids, or solids at room temperature.	Heavy isotopes behave just like other isotopes.	False	High
11	Buck did not like it, but he bore up well to the work, taking pride in it. It was a monotonous life, operating with machine-like regularity. One day was very like another. At a certain time each morning the cooks turned out, fires were built, and breakfast was eaten.	Buck was happy with his life despite its monotony.	False	Low

TABLE 6. *(Continued.)* **List of items.**

12	These days, neuroscience is beginning to catch up to musicians who practice mentally. Although the details are still somewhat elusive, the key to the success of mental imagery as a rehearsal technique is that most of the same neurological regions are invoked by mental practice as by real practice.	Mental practice of music is one of the myths that cannot yet be proven by neuroscience.	False	Medium
13	Hybrid vehicles have a halo that makes owners feel righteous and their neighbors feel guilty for not doing as much to save the planet. They also cost more and a new report from the very green Union of Concerned Scientists says buyers aren't always getting their money's worth.	Hybrid vehicles are a cost-effective way of contributing to lower carbon emissions.	False	Medium
14	The girl's feet were then re-wrapped even tighter than before, causing her footprint to shrink further as the bones slowly fused into their new configuration. Occasionally girls' feet would fester, and blood poisoning from gangrene could be a cause for concern, but an estimated 90% survived the process.	Blood poisoning and gangrene was a common cause of death for girls whose feet were reshaped for aesthetic reasons.	False	High
15	Very similar, but even more striking, is the evidence from athletic training. As with rehearsing a piece on the piano, practicing a complex physical task in the mind alone is nearly as effective a learning strategy as actually physically doing it. But it doesn't stop there.	Practicing mentally is nearly as effective as physical training when it comes to complex physical tasks.	True	Low
16	A bill was drafted and introduced into Parliament several times but met with great opposition, mostly from farmers. Eventually, in 1925, it was decided that summer time should begin on the day following the third Saturday in April and close after the first Saturday in October.	The introduced summer time was met with opposition.	True	Low
17	When early Europeans discovered Easter Island, its somewhat isolated ecosystem was suffering from the effects of limited natural resources, deforestation, and overpopulation. Over the following years the island's population of four thousand or so was slowly eroded by Western disease and deportation by slave traders.	Easter Island has been affected by overpopulation and disease.	True	High

TABLE 6. *(Continued.)* **List of items.**

18	Binge drinking may not necessarily kill or even damage brain cells, as commonly thought, a new animal study suggests. But it can block key receptors in the brain and trigger production of a steroid that interferes with brain functions critical to learning and memory.	Binge drinking can interfere with brain functions by triggering steroid production.	True	High
19	When attacked, a skunk's natural inclination is to turn around, lick its tail and spray a noxious scent. That works when a skunk faces a natural predator in the wild, but it's not as helpful when faced with, let's say, an oncoming car.	Skunks change their natural defense mechanism to deal with the peculiarities of human environments.	False	Medium
20	The astronauts used a hefty robotic arm to move the bus-size canister, stuffed with nearly three tons of packing foam and other space station refuse. It was the last job shared by the shuttle and station crews, numbering ten astronauts altogether.	The operation was performed by astronauts in both the shuttle and the station.	True	High
21	John Thornton asked little of man or nature. He was unafraid of the wild. With a handful of salt and a rifle he could plunge into the wilderness and fare wherever he pleased and as long as he pleased.	John was not afraid of going alone into the wilderness.	True	Medium
22	There often seems to be more diving in soccer than in the Summer Olympics. Phantom contact, or the slightest collision, can lead to theatrical belly flops and exaggerated somersaults by players deceptively trying to draw fouls on their opponents, kill time or catch a breather when tired.	Soccer players are similar to divers in terms of the skills they need in their respective fields.	False	High
23	Proper ventilation will make a backdraft less likely. Opening a room or building at the highest point allows heated gases and smoke to be released gradually. However, suddenly breaking a window or opening a door is a mistake, because it allows oxygen to rush in, causing an explosion.	In case of a fire, buildings should be opened at the highest point to avoid backdrafts.	True	Medium

TABLE 6. *(Continued.)* **List of items.**

24	Susan B. Anthony spent nearly sixty years of her life devoted to the cause of social justice and equality for all. Her major contributions were focused on women's rights. Her primary achievement lay in her inspiration and influence of thousands of people promoting the right of women to vote.	Susan B. Anthony's was a very influential advocate for social justice and gender equality.	True	Medium
25	As Earth warms due to the human-caused release of heat-trapping gases into the atmosphere, frozen Arctic soils also warm, thaw and release more carbon dioxide. The added carbon dioxide accelerates Earth's warming, which further accelerates the thawing of Arctic soils and the release of even more carbon dioxide.	Carbon dioxide generates a vicious cycle releasing even more carbon dioxide.	True	High
26	Unfortunately, for every six water bottles we use, only one makes it to the recycling bin. The rest are sent to landfills or, even worse, they end up as trash on the land and in rivers, lakes, and the ocean. Plastic bottles take many hundreds of years to disintegrate.	Plastic bottles stopped being a problem, because most of them are being recycled.	False	Low
27	For centuries, time was measured by the position of the sun with the use of sundials. Noon was recognized when the sun was the highest in the sky, and cities would set their clock by this apparent solar time, even though some cities would often be on a slightly different time.	Cities would synchronize their sundials with each other.	False	Low
28	American industry may not know it, but debate in Congress over reforming the federal government's regulatory apparatus may profoundly improve companies' ability to survive. This week the Senate seems poised to join the House in a massive overhaul of the way in which bureaucrats enact regulations that are strangling the life out of companies.	Collaboration between the bureaucrats and the senate is improving companies' ability to survive.	False	High
29	Stress is a risk factor for both depression and anxiety, he says. "We don't have data on the specific causes of depression and anxiety in this sample, but it does make sense scientifically that the Millennials who report higher levels of stress in their lives are also reporting higher levels of depression and anxiety."	Millenials with high levels of stress are also likely to suffer from depression and anxiety.	True	Low

TABLE 6. *(Continued.)* **List of items.**

30	When it comes to having a lasting and fulfilling relationship, common wisdom says that feeling close to your romantic partner is paramount. But a new study finds that it's not how close you feel that matters most, it's whether you are as close as you want to be, even if that's really not close at all.	The new study found that communication is what makes you feel closer to your partner.	False	Low

implementation of real-time subjective difficulty detection in a computerized learning system, with potential benefits for language learners and learners in general. Based on our results, it can be observed that generalization to new participants is particularly challenging. To compensate for this and make the system more useful, we recommend considering a fine-tuning step. Each new user could read and rate the difficulty of sample documents, and then feed these data into the model, improving its ability to predict subjective difficulty for each new user.

The second direction focuses on refining the models to achieve better performance. This could be accomplished through the identification and correction of the limitations of this work. As a first limitation, we used manually extracted features to train the machine learning models. While future research could explore a deep learning approach, such as that proposed by Reich et al. [4], there are also opportunities to improve the current methodology. For example, techniques such as Principal Component Analysis (PCA) and the exploration of nonlinear correlations with subjective difficulty could lead to useful insights regarding the selection of features for the classification models. A second limitation concerns the selection of the texts for the experiment. Although we selected the same number of texts with high, medium and low objective difficulty, only a quarter of the texts were rated as high difficulty by the participants. If the goal is to distinguish difficult texts from the rest (binary classification), future research should consider including more difficult texts in their experiments to provide a balanced class distribution, i.e., the same number of difficult texts as non-difficult texts. A third limitation is that we did not test all possible combinations of sensory modalities and features that might lead to better model performance. Given our research objectives, the only model variations we made (other than the full model) were to limit the physiological data to only one modality (hypothesis 1), and to shift the data (hypothesis 2). However, it is possible that certain combinations of modalities may lead to better results, compared to combining all modalities in the same model. In addition, all of our models included features that are relevant to language learning, i.e., language proficiency and objective text difficulty. However, we did not examine the effectiveness of these measures or the ability of physiological data alone to predict subjective text difficulty. Future research could explore these and other variations to better understand the factors that influence subjective text difficulty.

In conclusion, this work contributes to the literature on cognitive state detection by establishing a novel approach to predicting subjective text difficulty through a multimodal lens and by exploring the importance of data shifting to account for delayed physiological responses. It also contributes to the literature on language learning and L2 reading, by focusing on non-native speakers and including language proficiency, which is a measure that is particularly relevant to this group. The proposed directions for future research offer great potential for both practical applications and methodological advances. By integrating real-time subjective difficulty detection into computer-based learning systems, the possibilities for personalized and adaptive education could be greatly enhanced. At the same time, the pursuit of refining predictive models for subjective difficulty could be generalized to improve the detection of other cognitive states. Addressing the limitations identified in this work, including experimenting with other machine learning methods, text selection strategies, and different combinations of features, will be critical to these advances. As the fields of cognitive science and engineering continue to converge, this work serves as a foundation for further exploration and innovation, contributing to a broader understanding of cognitive processes and their applications.

APPENDIX

See Table 6.

REFERENCES

[1] R. Bahmani and M. T. Farvardin, "Effects of different text difficulty levels on Iranian EFL learners' foreign language reading motivation and reading comprehension," *Reading Foreign Lang.*, vol. 29, no. 2, pp. 185–202, 2017.

[2] G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, and M. Just, "User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention," *J. Med. Internet Res.*, vol. 15, no. 7, p. e144, Jul. 2013. [Online]. Available: http://www.jmir.org/2013/7/e144/

[3] C. S. Meppelink, E. G. Smit, B. M. Buurman, and J. C. M. van Weert, "Should we be afraid of simple messages? The effects of text difficulty and illustrations in people with low or high health literacy," *Health Commun.*, vol. 30, no. 12, pp. 1181–1189, Dec. 2015.

[4] D. R. Reich, P. Prasse, C. Tschirner, P. Haller, F. Goldhammer, and L. A. Jäger, "Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading," in *Proc. Symp. Eye Tracking Res. Appl.*, Jun. 2022, pp. 1–8.
[5] S. Ahn, C. Kelton, A. Balasubramanian, and G. Zelinsky, "Towards predicting reading comprehension from gaze behavior," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2020, pp. 1–5.
[6] N. Hollenstein, M. Tröndle, M. Plomecka, S. Kiegeland, Y. Özyurt, L. A. Jäger, and N. Langer, "Reading task classification using EEG and eye-tracking data," 2021, *arXiv:2112.06310*.
[7] S. Jacob, S. Ishimaru, and A. Dengel, "Interest detection while reading newspaper articles by utilizing a physiological sensing wristband," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 78–81.
[8] P. Vadiraja, J. Santhosh, H. Moulay, A. Dengel, and S. Ishimaru, "Effects of counting seconds in the mind while reading," in *Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2021, pp. 486–490, doi: 10.1145/3460418.3479357.
[9] A. P. Pai, J. Santhosh, and S. Ishimaru, "Real-time feedback on reader's engagement and emotion estimated by eye-tracking and physiological sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2022, pp. 97–98.
[10] P. Ayres, J. Y. Lee, F. Paas, and J. J. G. van Merriënboer, "The validity of physiological measures to identify differences in intrinsic cognitive load," *Frontiers Psychol.*, vol. 12, Sep. 2021, Art. no. 702538.
[11] U. Garain, O. Pandit, O. Augereau, A. Okoso, and K. Kise, "Identification of reader specific difficult words by analyzing eye gaze and document content," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 1346–1351.
[12] L. Tamor, "Subjective text difficulty: An alternative approach to defining the difficulty level of written text," *J. Reading Behav.*, vol. 13, no. 2, pp. 165–172, Jun. 1981.
[13] D. S. McNamara, A. C. Graesser, and M. M. Louwerse, "Sources of text difficulty: Across genres and grades," in *Measuring Up: Advances in How We Assess Reading Ability*. MD, USA: Rowman & Littlefield Education, 2012, pp. 89–116.
[14] R. Flesch, "A new readability yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, pp. 221–233, 1948, doi: 10.1037/h0057532.
[15] M. Brysbaert and B. New, "Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behav. Res. Methods*, vol. 41, no. 4, pp. 977–990, Nov. 2009.
[16] W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-U.K.: A new and improved word frequency database for British English," *Quart. J. Experim. Psychol.*, vol. 67, no. 6, pp. 1176–1190, Jun. 2014.
[17] E. Rodgers, J. V. D'Agostino, R. H. Kelly, and C. Mikita, "Oral reading accuracy: Findings and implications from recent research," *Reading Teacher*, vol. 72, no. 2, pp. 149–157, Sep. 2018.
[18] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, Jun. 1988.
[19] F. G. Paas, J. J. Van Merriënboer, and J. J. Adam, "Measurement of cognitive load in instructional research," *Perceptual Motor Skills*, vol. 79, no. 1, pp. 419–430, Aug. 1994.
[20] F. G. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach," *J. Educ. Psychol.*, vol. 84, no. 4, pp. 429–434, 1992.
[21] L. Fernandez, B. Höhle, J. Brock, and L. Nickels, "Investigating auditory processing of syntactic gaps with L2 speakers using pupillometry," *2nd Lang. Res.*, vol. 34, no. 2, pp. 201–227, Apr. 2018, doi: 10.1177/0267658317722386.
[22] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010.
[23] B. Schönebeck, K. W. Zimmer, and R. Kniesche, "Cognitive strain in text comprehension and heart rate variability," *Adv. Psychol.*, vol. 25, pp. 345–356, Jan. 1985.
[24] T. H. Khan, I. Villanueva, P. Vicioso, and J. Husman, "Exploring relationships between electrodermal activity, skin temperature, and performance during," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2019, pp. 1–5.
[25] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?" *Develop. Cogn. Neurosci.*, vol. 25, pp. 69–91, Jun. 2017.
[26] S. Ishimaru, S. S. Bukhari, C. Heisel, J. Kuhn, and A. Dengel, "Towards an intelligent textbook: Eye gaze based attention extraction on materials for learning and instruction in physics," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 1041–1045.
[27] S. Solhjoo, M. C. Haigney, E. McBee, J. J. G. van Merrienboer, L. Schuwirth, A. R. Artino, A. Battista, T. A. Ratcliffe, H. D. Lee, and S. J. Durning, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Sci. Rep.*, vol. 9, no. 1, p. 14668, Oct. 2019.
[28] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers Public Health*, vol. 5, p. 258, Sep. 2017.
[29] U. Nussinovitch, K. P. Elishkevitz, K. Katz, M. Nussinovitch, S. Segev, B. Volovitz, and N. Nussinovitch, "Reliability of ultra-short ECG indices for heart rate variability," *Ann. Noninvasive Electrocardiology*, vol. 16, no. 2, pp. 117–122, Apr. 2011.
[30] G. Geršak, "Electrodermal activity—A beginner's guide," *Electrotechnical Rev./Elektrotehniski Vestnik*, vol. 87, no. 4, pp. 175–182, 2020.
[31] Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt, and F. Vetere, "Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, Sep. 2017.
[32] J. Tromp, P. Hagoort, and A. S. Meyer, "Pupillometry reveals increased pupil size during indirect request comprehension," *Quart. J. Exp. Psychol.*, vol. 69, no. 6, pp. 1093–1108, Jun. 2016.
[33] D. Palomba, A. Angrilli, and A. Mini, "Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli," *Int. J. Psychophysiology*, vol. 27, no. 1, pp. 55–67, Jul. 1997.
[34] S. Sonkusare, D. Ahmedt-Aristizabal, M. J. Aburn, V. T. Nguyen, T. Pang, S. Frydman, S. Denman, C. Fookes, M. Breakspear, and C. C. Guo, "Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking," *Sci. Rep.*, vol. 9, no. 1, p. 4729, Mar. 2019.
[35] D. Leiner, A. Fahr, and H. Früh, "EDA positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure," *Commun. Methods Measures*, vol. 6, no. 4, pp. 237–250, Oct. 2012.
[36] C. D. Spielberger, H. F. O'Neil, and D. N. Hansen, "Anxiety, drive theory, and computer-assisted learning," *Prog. Exp. Personality Res.*, vol. 6, pp. 48–109, Aug. 1972.
[37] H. S. Nwana, "Intelligent tutoring systems: An overview," *Artif. Intell. Rev.*, vol. 4, no. 4, pp. 251–277, 1990.
[38] C. D. Le, "Using technology-enhanced language learning environments to influence the communicative potential of adult learners of English as a foreign language in Vietnam," Ph.D. dissertation, Inst. Sustain. Industries Liveable Cities (ISILC), Victoria Univ., Sydney, NSW, Australia, 2021.
[39] L. Canals and Y. Mor, "Towards a signature pedagogy for task-based technology-enhanced language learning: Design patterns," in *Proc. Eur. Conf. Pattern Lang. Programs*, Jul. 2020, pp. 1–11.
[40] F. M. Lord, "A broad-range tailored test of verbal ability," *ETS Res. Bull. Ser.*, vol. 1975, no. 1, pp. 1–12, 1975.
[41] P. Brusilovsky, "Methods and techniques of adaptive hypermedia," in *Adaptive Hypertext and Hypermedia*, 1998, pp. 1–43.
[42] M. E. Stetter and M. T. Hughes, "Computer-assisted instruction to enhance the reading comprehension of struggling readers: A review of the literature," *J. Special Educ. Technol.*, vol. 25, no. 4, pp. 1–16, Dec. 2010.
[43] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval, "Context-aware recommender systems for learning: A survey and future challenges," *IEEE Trans. Learn. Technol.*, vol. 5, no. 4, pp. 318–335, Oct. 2012.
[44] D. Di Mitri, J. Schneider, M. Specht, and H. Drachsler, "From signals to knowledge: A conceptual model for multimodal learning analytics," *J. Comput. Assist. Learn.*, vol. 34, no. 4, pp. 338–349, Aug. 2018.
[45] S. Ishimaru, N. Großmann, A. Dengel, K. Watanabe, Y. Arakawa, C. Heisel, P. Klein, and J. Kuhn, "HyperMind builder: Pervasive user interface to create intelligent interactive documents," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 357–360.
[46] K. Krishnasamy, "Code mixing among Tamil–English bilingual children," *Int. J. Social Sci. Humanity*, vol. 5, no. 9, pp. 788–792, 2015, doi: 10.7763/IJSSH.2015.V5.557.
[47] Empatica. *How is Ibi.csv Obtained?*. Accessed: Jul. 11, 2024. [Online]. Available: https://support.empatica.com/hc/en-us/articles/201912319-How-is-IBI-csv-obtained-

[48] Vercel Inc. *Next.JS*. Accessed: Jul. 11, 2024. [Online]. Available: https://nextjs.org
[49] S. G. Luke and K. Christianson, "The provo corpus: A large eye-tracking corpus with predictability norms," *Behav. Res. Methods*, vol. 50, no. 2, pp. 826–833, Apr. 2018.
[50] R. Speer, "Rspeer/wordfreq: V3.0," Zenodo, Version v3.0.2, Sep. 2022, doi: 10.5281/zenodo.7199437.
[51] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," *Behav. Res. Methods*, vol. 44, no. 2, pp. 325–343, Jun. 2012.
[52] J. A. E. Anderson, L. Mak, A. K. Chahi, and E. Bialystok, "The language and social background questionnaire: Assessing degree of bilingualism in a diverse population," *Behav. Res. Methods*, vol. 50, no. 1, pp. 250–263, Feb. 2018.
[53] Python Softw. Found. (2019). *Python Core Team, Python: A Dynamic, Open Source Programming Language*. [Online]. Available: https://www.python.org/
[54] W. McKinney, "Data structures for statistical computing in Python," in *Proc. Python Sci. Conf.*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61.
[55] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.
[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. J. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2012.
[57] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2008, pp. 2991–2996.
[58] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel, "Safe and sensible preprocessing and baseline correction of pupil-size data," *Behav. Res. Methods*, vol. 50, no. 1, pp. 94–106, Feb. 2018.
[59] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "CvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.
[60] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, Feb. 2021. [Online]. Available: https://link.springer.com/article/10.3758/s13428-020-01516-y
[61] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, "A method for stress detection using empatica E4 bracelet and machine-learning techniques," *Sensors*, vol. 23, no. 7, p. 3565, Mar. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3565
[62] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Feb. 2020.
[63] T. G. Dietterich, "Machine learning for sequential data: A review," in *Proc. Joint IAPR Int. Workshops Structural, Syntactic, Stat. Pattern Recognit.*, Windsor, ON, Canada. Cham, Switzerland: Springer, Jan. 2002, pp. 15–30.
[64] R. Pandey, S. K. Khatri, N. K. Singh, and P. Verma, *Artificial Intelligence and Machine Learning for EDGE Computing*. New York, NY, USA: Academic, 2022.
[65] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Jul. 2001.
[66] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler, "Validation of the empatica E4 wristband," in *Proc. IEEE EMBS Int. Student Conf. (ISC)*, May 2016, pp. 1–4.
[67] A. A. T. Schuurmans, P. de Looff, K. S. Nijhof, C. Rosada, R. H. J. Scholte, A. Popma, and R. Otten, "Validity of the empatica E4 wristband to measure heart rate variability (HRV) parameters: A comparison to electrocardiography (ECG)," *J. Med. Syst.*, vol. 44, no. 11, pp. 1–11, Nov. 2020.
[68] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, Oct. 2018, pp. 400–408, doi: 10.1145/3242969.3242985.

JAVIER MELO was born in Santiago, Chile, in 1993. He received the B.S. and M.S. degrees in industrial and organizational psychology from the Pontifical Catholic University of Chile, in 2018, and the M.S. degree in cognitive science from the University of Kaiserslautern-Landau, Germany, in 2023.

From 2018 to 2019, he worked in human resources consulting companies, doing people analytics and software development. Since 2023, he has been a Researcher with German Research Center for Artificial Intelligence (DFKI), Germany. His research interests include artificial general intelligence, technology-enhanced learning, and human–computer interaction.

LEIGH FERNANDEZ received the B.Sc. degree in psychology and the B.A. degree in primate behavior and ecology from Central Washington University, Ellensburg, WA, USA, in 2005, the M.Sc. degree in clinical linguistics from the University of Potsdam, Potsdam, Germany, in 2009, the M.Phil. degree in psychology from Northumbria University, Newcastle, U.K., in 2014, and the Ph.D. degree in brain and language from the University of Potsdam, in 2016.

Since 2015, she has been a Senior Research Scientist with the University of Kaiserslautern-Landau. Her research interest includes language processing (both written and spoken) in first and second language speakers across the adult life-span.

SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. and M.E. degrees in electrical engineering and information science from Osaka Prefecture University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree (summa cum laude) in engineering from the University of Kaiserslautern, Germany, in 2019.

He has been a Project Professor with the Department of Computer Science, Osaka Metropolitan University, Japan, since 2023. In addition, he has been a Researcher with the Keio Media Design Research Institute, since 2014. He was a Junior Professor with the University of Kaiserslautern-Landau, from 2021 to 2023, and has been a Senior Researcher with German Research Center for Artificial Intelligence (DFKI), Germany, since 2019. His research interests include human–computer interaction, machine learning, and cognitive psychology with the aim of amplifying human intelligence.

Prof. Ishimaru's awards and honors include Best Presentation Award at Asian CHI Symposium, in 2020, Poster Track Honorable Mention at UbiComp/ISWC, in 2018, and MITOU Super Creator which is a title given to outstanding software developers (around ten people per year) by the Ministry of Economy, Trade, and Industry in Japan.