scientific reports

Check for updates

OPEN

Using eye movements, electrodermal activities, and heart rates to predict different types of cognitive load during reading with background music

Ying Que[1], Yueyuan Zheng[3], Janet H. Hsiao[3,4] & Xiao Hu[1,2]✉

The triarchic model of cognitive load postulates three types of cognitive load—extraneous, intrinsic, and germane load. While various approaches have been proposed to measure the three types of cognitive load, most measurements are intrusive. To address this issue, we leveraged multimodal learning analytics to collect eye movement (EM), electrodermal activity (EDA), heart rate (HR), and heart rate variability (HRV) from non-intrusive sensors and investigate whether they could predict the three types of cognitive load. We examined extraneous load (created by adding background music (BGM)), intrinsic load (created by text complexity), and germane load (reflected by comprehension accuracy) in a novel reading context with self-selected preferred BGM. One hundred and two (102) non-native English speakers were recruited. Half of them read English passages with BGM, while the other half read in silence. Results of logistic regression indicated that EM measures were predictive of the three load types, while HR/HRV measures predicted extraneous and germane load. Our findings provide evidence supporting the triarchic structure of cognitive load theory and implications for the design of non-intrusive measurement of cognitive load.

Keywords Cognitive load, Reading comprehension, Background music, Eye movement, Electrodermal activity, Heart rate

Computer-assisted learning technology offers the potential to enhance tailored learning experiences for learners worldwide. To realize this potential, instructional systems need to adapt to individual learners' personalized learning processes[1]. With recent development of wearable computing and multimodal learning analytics (MmLA)[2], one possible approach to developing such adaptable systems is to use learners' real-time eye movements (EM), and peripheral physiological signals, such as electrodermal activity (EDA), heart rates (HR), and heart rate variability (HRV) to capture and monitor their cognitive states during learning[1,3].

Learners' cognitive states can be influenced by three types of cognitive load: extraneous, intrinsic, and germane load, which can either facilitate or impede the learning process[4,5]. Sweller's cognitive load theory[6] introduced fundamental concepts of the three types of cognitive load, and since then there have been many studies that measured the three types of cognitive load[1,4]. However, most of the extant literature relies on learners' self-reports as measurements of their cognitive load[4,7], which can be distractive and/or inaccurate. If administered during learning, they require learners to pause the ongoing tasks to report their cognitive processes; if administered post-learning, they would fail to capture moment-to-moment variations in cognitive load[1].

Inspired by recent development of wearable computing and its application to research on learning[8], we set out to explore unobtrusive methods for measuring cognitive load. In particular, we are interested in a novel reading context with self-selected preferred background music (BGM). Music permeates many aspects of human society, accompanying activities related to daily study and work. Previous studies have revealed mixed findings regarding how BGM impacts reading, leaving the question of whether BGM can promote or hinder reading still open[9–11]. Some studies observed positive effects, such as mood adjustment or arousal elevation[9–12], whereas

[1]Faculty of Education, The University of Hong Kong, Hong Kong, China. [2]College of Information Science, The University of Arizona, Tucson, AZ, USA. [3]Division of Social Science, Hong Kong University of Science and Technology, Hong Kong, China. [4]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China. ✉email: xiaohu@arizona.edu

others discovered negative effects, such as cognitive overload or distraction[9,10,12–14]. It can be seen that cognitive load plays an important role here. This study thus aims to provide empirical evidence regarding whether and the extent to which physiological signals such as EM, EDA, HR/HRV can relate to different types of cognitive load, while also advancing scientific understanding of the underlying mechanisms about BGM.

Literature review

Theoretical framework for cognitive load

Meaningful learning requires a significant investment of cognitive resources. Cognitive load theory (CLT) can be used to understand learners' allocation of cognitive resources and provide guidance for instructional design[5]. CLT is based on two assumptions: (1) each individual's working memory capacity (WMC) is finite, and (2) schema-related mechanisms exist. Regarding the first assumption, CLT assumes that human memory is made up of a finite amount of WMC and an infinite amount of long-term memory[15]. If the cognitive load imposed by the learning task exceeds WMC, the learning process would be impeded[16]. As for the second assumption, schema refers to knowledge structures organized around core concepts[17]. It enables learners to categorize information based on how it will be used. Schema acquisition and rule automation are two primary mechanisms of learning, both of which make use of long-term memory to compensate for the lack of WMC[1]. During the learning process, learners use schemas stored in long-term memory and integrate them with new knowledge, resulting in an increase in both the quantity and quality of schemas[1,17]. CLT provides theoretical guidelines for this study in investigating learners' cognitive load when they read with self-selected preferred BGM or in silence.

Extraneous cognitive load pertains to unproductive cognitive processes that use working memory resources to handle irrelevant elements that are not related to the learning objectives and do not contribute to the learning outcomes (e.g., schema acquisition or automation). It is typically determined by the way instructional information is presented. Poor instructional design can increase the extraneous load, making it harder for learners to understand the elemental interactions necessary for learning given materials[5,6]. According to CLT, extraneous load should be avoided in instructional design to prevent interference with learning.

Intrinsic cognitive load is closely linked to the cognitive processes that are essential for learners to comprehend the learning material, which depends on the complexity of the material, namely the number of interacting elements that need to be processed in working memory at any one time. Intrinsic load can differ depending on learners' prior knowledge[1]. A learner with a higher level of prior knowledge may find the same learning material less challenging than a learner with a lower level of prior knowledge. When learners are presented with materials of higher complexity, they will experience a greater intrinsic cognitive load. Indeed, previous research has found that, the cognitive load required to maintain the activity across multiple elements increases with the complexity of the learning materials[6].

Germane cognitive load refers to the cognitive effort used in effective learning, understanding, and constructing new schemas (i.e., knowledge structures organized around core concepts)[17]. In other words, it is the load that contributes to organizing information, establishing connections between new information and learners' pre-existing knowledge, as well as constructing new knowledge in long-term memory, and it demonstrates a positive connection to learning outcomes[1,4,5]. Ayres suggested that effective instructional design should increase germane load while minimizing extraneous load[18].

CLT suggests that cognitive processes are limited by the capacity of working memory[5,15]. Individuals' WMC varies from person to person, which can affect the extent to which cognitive load is managed effectively[19]. An experimental study[19] showed that individuals with higher WMC could better handle attentional and memory tasks and manage multiple sources of information simultaneously, whereas individuals with lower WMC might struggle to process and maintain the same amount of information. Therefore, in order to accurately measure the effects of examined cognitive load factors on learning outcomes, it is important to consider individual differences in WMC.

Individuals' skill level and prior knowledge can influence how much cognitive load is experienced during a task[1]. When individuals have a high level of skill or prior knowledge in a particular domain, they are able to process information more efficiently, which reduces the cognitive load associated with that task[5]. When it comes to the English as Second Language (ESL) reading comprehension task, learners' English proficiency can serve as an indicator of their skill level and prior knowledge in language. English proficiency was identified as one of the strongest predictors of reading comprehension performance in ESL learners[20]. ESL learners who had a larger English vocabulary size, which is an indicator of greater language proficiency, were found to have scored higher in reading comprehension tasks[21]. Given that participants in this study were all ESL learners and their task was to read English passages, individual differences in English proficiency should be considered as they may affect cognitive load during reading.

Multimodal learning analytics for cognitive load measurement

In the interdisciplinary field of learning analytics, researchers advocate for gathering multimodal data to examine learners' behaviours, cognition, and affect, including measuring cognitive load and its influence on academic achievement or other learning related constructs[2]. This approach is known as multimodal learning analytics (MmLA), which leverages subjective and objective methods to capture learners' cognitive processes.

Psychometric assessments have been widely used to measure cognitive load due to their ease of administration, high reproducibility, and sensitivity to subtle changes in task demands[7]. These scales are often the gold standard for evaluating cognitive effort which other types of assessments are compared against[22,23]. The underlying assumption of using subjective rating scales is that participants can reflect on their thoughts and feelings and report the degree of mental effort they expended in the precedent task[7]. NASA Task Load Index (TLX) index and 9-point Paas scale are two commonly used psychometric rating scales for evaluating cognitive load[22,24]. However, subjective ratings are usually presented after learning activities, which may not be able to provide

continuous information about the ongoing cognitive load[1]. On the other hand, repeated rating during learning may bias the results of cognitive load measurements due to the frequent disruption of learning activities and fatigue of learners[25].

The use of biological indicators (e.g., eye-tracking and peripheral physiological signals) to measure cognitive load can be effective owing to the continuous nature of the measurement involved, and the method is capable of generating highly detailed data and tracking cognitive load continuously throughout the learning process[3,10]. Among biological indicators, eye movements (EM) measures have garnered interest among researchers studying cognitive load[1,10]. EM measures commonly used in these studies include those related to fixations such as fixation duration (i.e., the amount of time eyes remain relatively stationary), and those related to saccades such as saccade amplitude (i.e., the distance between two fixations), as well as regressive EM (i.e., rereading of previously fixed words) such as regression count. Researchers have used these measures to assess cognitive load in reading[10,26,27]. For instance, fixation duration and saccade amplitude can reflect cognitive load in lexical processing, while regressive EM can indicate difficulties in post-lexical semantic integration[27].

Electrodermal Activity (EDA) refers to alterations in electrical properties of the skin that are mediated by the level of physiologically induced sweating[28]. The phasic components of the EDA signal are responsible for rapid fluctuations or peaks and are closely tied to the onset of a stimulus[29]. Skin Conductance Response (SCR) is the most common measure of the phasic components, which can be triggered by emotion and stress leading to the activation of the sympathetic nervous system's sudomotor nerves[29]. Fluctuations in SCR events have been used to indicate the most stressful or cognitively demanding moments in the evaluation of multimodal interfaces[30] or sudden fluctuations in cognitive load among physicians[31].

Numerous studies have demonstrated a direct relationship between cognitive task demands and variations in cardiac activity[32–34]. However, to the best of our knowledge, only a few published studies have investigated such relationships in the context of reading comprehension. Scrimin et al.[34] examined heart rate (HR) and heart rate variability (HRV) as physiological measures while students read a science text. The findings showed a general trend that students' HRV was higher during the reading phase compared to the baseline, but their HR decreased during the reading process. This pattern suggests that students exert deep cognitive effort while engaging in reading comprehension. Daley et al.[32] investigated cardiac vagal tone in middle school students during a reading task, where they were asked to read two passages and recall them orally. The results revealed that respiratory sinus arrhythmia, the tested cardiovascular indicator of physiological responses, predicted reading comprehension as indicated by their verbal retelling performance.

Research gaps and research questions

Based on the related work, we identified several research gaps. First, whereas many studies have employed subjective surveys as cognitive load measures, few have incorporated multimodal measures through the implementation of physiological sensors[1,29,31]. Second, to the best of our knowledge, only a few published studies[1,4] have examined Sweller's three types of cognitive load (i.e., extraneous, intrinsic, germane load)[6] simultaneously and then investigated the connections between the types of cognitive load and potential cognitive load indicators including eye-tracking metrics and self-reported mental efforts ratings. In particular, there is a lack of research that examines the connection between EDA, and HR/HRV metrics with three types of cognitive load.

This study aimed to investigate the feasibility of using EM, EDA, HR/HRV from non-intrusive sensors to measure the three types of cognitive load[5], while taking learner characteristics into account. Specifically, this study focuses on the following research question (RQ) in the reading context with and without self-selected BGM.

RQ: To what extent can multimodal sensor data and learner characteristics predict the three hypothesized cognitive load types (i.e., extraneous, intrinsic, germane load)?

Methods

The experiment was approved by The University of Hong Kong's Human Research Ethics Committee (reference number: EA1802092). All methods were carried out in compliance with the American Psychological Association ethical standards. All participants gave their informed consent prior to their involvement in the user experiment.

Participants recruitment requirement

Participants should be English as a Second Language (ESL) learners aged between 18 and 35, without any visual, hearing, or learning impairments. A priori power analysis was conducted using G*Power[35], informed by empirical effect sizes reported in DeLeeuw and Mayer[4], as well as guidelines referenced in Zu et al.[1]. The analysis considered anticipated effects across different types of cognitive load (extraneous, intrinsic, and germane). Based on these estimations, a sample size of up to 94 participants was considered sufficient to detect the effect sizes reported in prior literature[4] with 80% power at $\alpha = .05$. Our final sample included 102 participants, which exceeded this threshold. Detailed sample size calculations are provided in the Appendix in the supplementary material. Generally, eye movement studies analyzing the effects of manipulated cognitive factors on eye-tracking measures consider a sample size exceeding 50 participants to be large[1].

Procedure

As Fig. 1 shows, participants first filled in a pre-survey that gathered their demographic information (e.g., gender, age, major). They then took the LexTale test, which measures general English proficiency. Stabilization on the Empatica E4 wristband was administered to ensure the quality of peripheral physiological signals. After that, participants were asked to sit at total rest for 2-min to record the baselines of peripheral physiological signals. The facilitator guided participants through a practice block to familiarize them with the formal experiment.

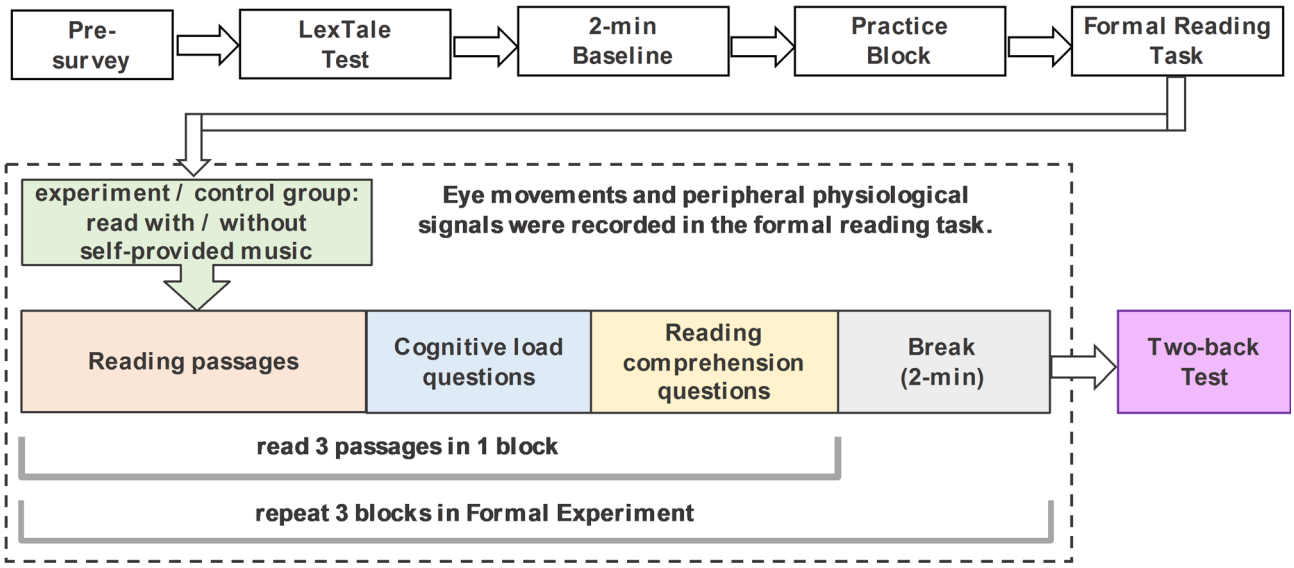

Fig. 1. Experimental procedure.

The main task required participants to read all the passages either with BGM (for those in the experimental group) or in silence (for those in the control group), during which the movements of their dominant eye were tracked and their peripheral physiological signals were recorded. The nine passages were evenly assigned into three blocks. The order of the blocks and passages was counter-balanced with regard to difficulty levels (see below) across participants. To ensure the accuracy of eye tracking, eye-tracker calibration was performed before each block, and drift correction was conducted before each passage. After finishing a passage, the participants were instructed to press the "continue" button, and answered two cognitive load questions regarding perceived difficulty level and extent of understanding of the passage in sequence, followed by two reading comprehension questions on the content of the passages. Participants were given unlimited time for reading and answering questions, which was to simulate a typical reading context instead of examinations. After each block, participants had a 2-min rest to relieve potential tiredness. After three blocks, they then completed the two-back test, which measures WMC.

Reading task

The reading task was composed of nine English passages which were selected from the reading comprehension samples from Graduate Record Examinations (GRE) with a variety of general topics. The passages were divided into three levels of text difficulty determined by a widely used text readability score, Flesch-Kincaid grade[36]. The higher the grade, the harder it is to read a text. We calculated this metric using the online software readable. io. Means and standard deviations (in parenthesis) of the grade of passages in each level were: 13.3 (0.35) for easy, 17.4 (0.46) and 21.0 for hard (0.47). The easy-level passages covered themes of astronomy, astrobiology, and physics; the medium-level passages covered themes of archaeology, history, and literature; the hard-level passages covered themes of biology, sociology, and anthropology.

The passages covered a diversified scope of themes including astronomy, astrobiology, physics, archaeology, history, literature, biology, sociology, and anthropology, aimed at mitigating potential biases arising from learners' varied knowledge backgrounds. All passages expressed complete ideas or meanings and contained a similar number of words, with an average word count of 218 and a standard deviation of 9.8. For each passage, we designed one text-based question, which was conceptually simple and only required shallow understanding of the content, and one inference-based question, which required reasoning and deep understanding of the content[10,37]. Each question had four response alternatives. Accuracy in answering the text-based and inference questions were assessed.

Self-reported cognitive load questions

The questions were used to evaluate students' self-perceived difficulty level and extent of understanding of the passage. For self-perceived difficulty of the passage, we asked "How difficult do you think the passage is?", which was adapted from the NASA Task Load Index (TLX) Mental Demand question[22]. For self-perceived understanding of the passage, we asked "To what extent do you understand the passage?", which was adapted from the TLX Performance question[22]. Each question was rated and recorded on a 5-point Likert scale, with 1 being the lowest rating (i.e., very easy, not understanding at all) and 5 being the highest rating (i.e., very difficult, understanding very well).

LexTale test

English proficiency was assessed by the LexTale test that could measure participants' familiarity level to English words and represent their general English proficiency[38]. The test contained 60 trials and in each trial, a string of

letters was presented. Participants judged if the string was an existent English word or not. English proficiency was computed as the percentage of correct responses in the LexTale test.

Two-back test

WMC was examined by two-back tests[39]. In the tests, a continuous stream of single letters was shown at different locations one at a time. Participants determined whether the current letter (in the verbal subtask) or the current location (in the spatial subtask) was the same as the one presented two trials earlier. Each subtask consisted of 36 trials, each letter lasting 1000 millisecond (ms) followed by a 2500 ms blank screen. WMC was measured by average accuracy of all trials across verbal and spatial subtasks.

Apparatus

EyeLink 1000 plus (tower mount model) was used to record participants' eye movements during reading the passages. The eye tracker's sampling rate was set as 2000 Hz, and resolution of the computer monitor (19 inches) was 1280 * 1024 pixels for displaying reading passages. The viewing distance was 56 cm, and horizontal visual angle for each English character was around 0.3°, which simulated a normal reading situation[40]. A research-grade wristband, Empatica E4 was worn by the participants to record peripheral physiological signals including Heart Rate (HR), Inter-beat Intervals (IBI), Electrodermal Activity (EDA), etc. All recorded data were synchronized by timestamps and were anonymized for keeping confidentiality.

Design for examining three types of cognitive load

We examined three types of cognitive load following the experimental design adopted in previous studies[1,4].

Extraneous load was created based on the coherence principle which proposed that irrelevant information should be avoided in multimedia presentations[16]. According to the coherence principle, irrelevant sound such as BGM is not suggested to be presented along with the text because it would require cognitive resources to either attend to or to inhibit the irrelevant information[16]. In the present study, participants were randomly assigned to either BGM or silence condition. In the BGM condition, passages were presented with the accompaniment of participants' self-provided preferred BGM. Since participants needed to either process or inhibit the irrelevant sound while reading, listening to BGM was supposed to increase their extraneous cognitive load, as reflected in the impaired performance on reading comprehension tests in previous studies[13,14]. The background audio condition (i.e., BGM vs. silence) was thereby a between-subject factor.

Intrinsic load depends on the complexity of the learning material and was created with text complexity determined by Flesch-Kincaid Grade[36] (as described in Reading Task). Based on the grades, passages were evenly coded into Easy, Medium, and Hard levels. The level of passage complexity was therefore a within-subject factor in the experiment.

Germane load is defined as the cognitive resources that are available to process the intrinsic load[1,4]. It is linked to effective learning and is indicative of deep cognitive processing, as demonstrated by activities such as mentally organizing learning materials. This type of cognitive load is instrumental in enhancing learners' performance[18]. In the present study, germane load was reflected using reading comprehension test scores. This assumes that those who comprehend materials better would have invested more cognitive resources to process the material more elaborately, namely, to have experienced higher germane load, and those who experienced high or low germane load during the reading phase would score higher or lower in the reading comprehension tests. Similar to DeLeeuw and Mayer's study on multimedia lessons[4], we divided participants into low and high germane load groups using a mean split on the passage's reading comprehension score. That is, the participants who had a better-than-average score were designated as having experienced high germane load, and those who had a lower-than-average score were designated as having experienced low germane load. The reading comprehension test score was thus a between-subject factor.

Proposed measures

This study collected questionnaire, EM, EDA, and HR/HRV data during the reading task.

Questionnaire

Self-reported measures were evaluated by two subjective questions which asked the participants' perceived difficulty level, and understanding degree of the passage (see the exact questions in Self-reported Cognitive Load Questions). The ratings on the two scales were analyzed separately, each of which was equal to the averaged value across passages.

Eye movement metrics

Table 1 shows eye movement (EM) features we used to infer readers' cognitive processes, including mean fixation duration, mean saccade amplitude, fixation count, and regression count[26]. Longer and more fixations can indicate a task that is more cognitively demanding, and the individual is experiencing cognitive overload[41]. Saccade amplitude indicates the length of rapid movements between two fixations. Individuals with reading problems often exhibit shorter saccade amplitudes[26], and a reading task with high cognitive demand would result in shorter saccade amplitude[27]. Regressive EM behaviours involve moving backwards to the word previously being fixated on. A cognitively demanding reading task would typically cause increased regressions[27,42]. Mean Fixation Duration and Mean Saccade Amplitude are measures at word-level, aggregated as mean values across all words within a passage. Fixation Count and Regression Count are measures at passage-level, aggregated as sum values per passage.

Measures	Descriptions
Mean fixation duration	Mean length of time spent on fixating on a word within a passage
Mean saccade amplitude	Mean amplitude of all saccades within a passage
Fixation count	Total number of fixations within a passage
Regression count	Total number of regressions within a passage

Table 1. Eye movement measures and their descriptions.

Signals	Measures	Descriptions
EDA	SCR frequency	Total number of the detected SCR peak per second
	SCR amplitude	Mean of the amplitude of the detected SCR peak
HR/HRV	Mean HR	Number of heart beats per minute
	HRV RMSSD	Root mean square of successive differences between normal heartbeats

Table 2. Peripheral physiological measures and their descriptions.

Electrodermal activity, heart rate, and heart rate variability metrics

Table 2 shows peripheral physiological measures and their descriptions. Electrodermal Activity (EDA) refers to the signals produced by sweat on the skin. Skin Conductance Responses (SCRs) refer to the phasic activity of EDA signals, which fluctuate rapidly. SCRs are innervated by sudomotor nerves of the sympathetic nervous system, firing in response to emotional and stressful stimuli[29]. Therefore, fluctuations in SCR are thought to be salient indicators of changes in the extent of stress (e.g., cognitive load) induced by a stimulus or event[43]. Similar to prior studies[31,44], we calculated SCR Frequency (i.e., the total number of the detected SCR peak per second) and SCR Amplitude (i.e., the mean of the amplitude of the detected SCR peak) as potential indicators of cognitive load.

Heart Rate (HR) and Heart Rate Variability (HRV) are two main cardiac response constructs commonly studied in cognitive load research. As cognitive demand increases, HR rises while HRV decreases[33]. This is because during psychophysiologic arousal state, heart rhythm becomes faster and more uniform. In this study, we chose the mean HR (beats per minute), and one widely used and psychologically meaningful HRV feature, RMSSD (i.e., the root mean square of successive differences between normal heartbeats). RMSSD computes the differences between adjacent heartbeats in milliseconds, squares these values, and the result is averaged before taking the square root of the total; it can reflect larger changes from one beat to the next[45,46].

Preprocessing eye movements and peripheral physiological signals

Automatic parser of Eyelink 1000 plus with default settings for cognitive research was used to identify saccades and fixations: if an eye movement had an instantaneous velocity greater than 30°/sec or an acceleration greater than 8000°/s^2, it was categorized as a saccade, and the remaining data points between successive saccades were classified as fixations. The Eyelink DataViewer was then used to generate a series of eye movement data. Outlying eye movements, such as those beyond the image stimuli and those caused by drift correction, were removed.

The difference in sampling rates were considered (EDA: 4Hz; HR: 1Hz) when pre-processing different kinds of peripheral physiological signals. All the signals were segmented based on the start timestamp and end timestamp of the period spent in reading each passage or the 2-min rest period at the beginning of the experiment that measured the baseline of the peripheral physiological signals. Features were further derived from the segments according to the corresponding feature extraction methods as follows.

Analysis of raw EDA signals was performed with neurokit2, a Python-based toolkit, that can detect peaks of the EDA signals, namely the Skin Conductance Response (SCR) events[47]. In pre-processing, we first filtered noises in the EDA signals through a unidirectional first-order Butterworth low pass filter with a cut-off frequency of 0.05 Hz[48] and then decomposed the signals into the phasic and tonic component by using the cvxEDA method[28] which was provided by the eda_phasic function of neurokit2 (with a sampling rate of 4 Hz). After that, we detected SCR peaks from the phasic component with the eda_peak function before computing frequency and amplitude of SCR.

We used hrvanalysis, a Python package for heart rate variability analysis[49] to compute the HRV RMSSD feature from the IBI data provided by Empatica E4 (see how to compute it in the Electrodermal Activity, Heart Rate, and Heart Rate Variability Metrics section). In pre-processing the signals, we first adopted hrvanalysis package to detect outliers from RR intervals above 2000 or below 300 milliseconds, as well as identify ectopic beats via the Malik method: Intervals deviating by greater than 20% from the preceding interval were regarded as outliers and replaced with linear interpolation[45]. Moreover, we computed the mean HR based on the HR data directly provided by Empatica E4.

Data analysis

To minimize the potential impact of individual differences across the two background audio conditions (i.e., BGM vs. silence), we applied independent sample t-tests to compare two learner characteristics, namely English proficiency (measured by LexTale test scores) and WMC (measured by two-back test accuracy), as well as the

baseline measurements of the peripheral physiological signals from 2-min resting periods between the two conditions.

Before answering the RQ, we performed Mann-Whitney U tests (for ordinal variables) to check if there was any difference in each self-reported measure between various conditions (i.e., BGM vs. silence or low vs. high comprehension accuracy, based on between-subject design). Moreover, one-way repeated measures ANOVA was conducted to examine if there was any difference in each self-reported measure across intrinsic load conditions (i.e., easy vs. medium vs. hard text complexity, based on within-subject design). Mauchly's test was used to assess data sphericity. When data violated sphericity, degrees of freedom were corrected using Huynh-Feldt (estimated epsilon $\varepsilon > 0.75$) or Greenhouse-Geisser correction ($\varepsilon < 0.75$)[50]. Significant main effects were followed with Bonferroni Post-hoc for three of the paired samples comparison (corrected $p = 0.016$).

To answer the RQ, we constructed binary logistic regression models (method: stepwise) to predict cognitive load types (i.e., conditions of extraneous, intrinsic, germane load serve as dependent variables) by simultaneously including multimodal sensor data and learner characteristics as independent variables. Measures of multimodal sensor data were extracted from EM, EDA, and HR/HRV signals that were collected when participants were reading passages. Learner characteristics included WMC and general English proficiency. Only statistically significant models with significant predictors were reported. We performed a series of likelihood ratio tests to evaluate the overall goodness of fit of the logistic regression models, denoted by chi-square statistics, and the associated p-values, complemented by R^2 measures[51]. In particular, the Nagelkerke's R^2 was used to represent how well the independent variables explain the variance in the dependent variable in a logistic regression model.

Results

Participants

One hundred and two (102) participants (52 females, 50 males) in The University of Hong Kong were recruited. The participants majored in diversified fields, such as engineering, computer science, mathematics, education, psychology, geography, pharmacy, etc. All of them were ESL learners between 18 and 35 years old (Mean = 23.68, SD = 4.22). All participants reported with normal or corrected normal eyesight and without visual, hearing, language, or learning impairment. They were randomly and evenly assigned to the BGM group (51) and the silence group (51) to read English passages with their self-provided BGM or in silence. Eye movement data from one participant (one in the silence group) and peripheral physiological data from six participants (four in the BGM group and two in the silence group) were excluded from subsequent data analysis due to device malfunction.

Benchmarking individual differences in two background audio conditions

To assess whether learners' English proficiency, WMC, and baselines of peripheral physiological signals confounded with the effects of background audio conditions (i.e., BGM vs. silence) on cognitive load, independent sample t-tests were conducted on participants' LexTale test scores, two-back test accuracy, and baselines of peripheral physiological features across the two background audio conditions, yielding no significant differences in score of LexTale test: $t(100) = .358$, $p = .721$; accuracy of two-back test: $t(100) = .708$, $p = .481$; baselines of SCR Frequency: $t(94) = -1.779$, $p = .078$; SCR Amplitude: $t(78) = 1.403$, $p = .164$; Mean HR: $t(94) = 1.131$, $p = .261$; and HRV RMSSD: $t(94) = .261$, $p = .794$.

Descriptives of BGM characteristics

Previous studies have shown that the types of BGM[52], tempo[53], presence of lyrics[54], and languages of lyrics[55] in the BGM can impact reading comprehension. Therefore, we analyzed and reported the distributions of these four characteristics in the BGM playlists provided by participants in the BGM group (N = 51). Music genres, presence of lyrics, and languages of the lyrics were coded manually. Specifically, the music genres were classified based on an existing genre taxonomy[56]. Two researchers first independently coded the BGM pieces, followed by consensus discussion and majority voting by a third researcher when discrepancies remained. Tempo was extracted with Librosa[57], a well-known music audio processing library, and categorized into slow, moderate, and fast based on extant literature[58]. Results showed that the most common musical genre was rhythmic and intensity (e.g., pop, hip-hop; 54.9%), followed by classical music (e.g., piano, organ; 23.5%), rebellious genres (e.g., heavy metal, punk; 7.8%), easy listening (e.g., country, folk; 5.9%), electronic (2.0%), jazz & blues (2.0%), and mix genres (3.9%). Regarding lyrics presence, 61% included lyrics, 31% were instrumental, and 8% contained both lyrical and instrumental music. Average tempo was 134.89 BPM (SD = 26.72), with most playlists featuring moderate tempos (88.2%), and a few with fast (5.9%) and slow tempos (5.9%). Languages of lyrics included English (42.9%), Mandarin (31.4%), Cantonese (28.6%), Japanese (17.1%), Korean (11.4%), Hindi (2.9%), and Polish (2.9%).

Effects of three hypothesized types of cognitive load on self-reported measures

We compared each self-reported measure between different conditions.

First, we compared measures between two background audio conditions which differed in assumed extraneous load. Results showed that there were no significant differences in self-reported levels of passage difficulty ($U = 1283.50$, $p = .912$, $r = .013$) or passage understanding ($U = 1423.00$, $p = .413$, $r = -.094$) between the BGM and silence condition. Participants in the BGM condition reported comparable levels of passage difficulty (Mean = 2.90, SD = .57) and passage understanding (Mean = 3.20, SD = .54) to those reported in the silence condition (passage difficulty: Mean = 2.85, SD = .67; passage understanding: Mean = 3.28, SD = .63). The lack of significant differences in self-reported levels of passage difficulty and passage understanding between the BGM and silence conditions may be attributed to the use of self-selected BGM. Allowing participants to choose

Targets	Predictors	B (S.E.)	OR [95% CI]	Sig	R^2
Extraneous: Silence (0) vs. BGM (1)	Fixation count	0.002 (0.000)***	1.002 [1.001, 1.003]	0.000	0.082***
	Mean HR	− 0.034 (0.007)***	0.966 [0.954, 0.979]	0.000	
	Constant	1.925 (0.510)***	6.858 [NA, NA]	0.000	
Intrinsic: Easy (0) vs. Hard (1)	Mean saccade amplitude	0.454 (0.099)***	1.575 [1.296, 1.913]	0.000	0.072***
	Fixation Count	0.002 (0.001)***	1.002 [1.001, 1.003]	0.000	
	Constant	− 3.070 (0.608)***	0.046 [NA, NA]	0.000	
Germane: Low (0) vs. High (1)	Fixation Count	0.001 (0.000)**	1.001 [1.000, 1.002]	0.002	0.021**
	HRV RMSSD	0.008 (0.004)*	1.008 [1.001, 1.015]	0.033	
	Constant	− 0.761 (0.264)**	0.467 [NA, NA]	0.004	

Table 3. Logistic regression analyses in predicting conditions of cognitive load types. B = coefficients, S.E. = standard errors associated with the coefficients. OR = Odds Ratios for the predictors that were calculated by Exp (B), R^2 as computed with Nagelkerke R^2. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

their own BGM could introduce variability in the impact of the music, contributing to inconsistent effects on perceived cognitive load.

Second, we conducted one-way repeated measures ANOVA to compare each proposed measure across easy, medium, and hard levels of text complexity which mark different intrinsic load conditions. Salient main effects of text complexity were observed on self-reported passage difficulty level ($F(2,202) = 54.975$, $p < .001$, $\eta^2 = .352$) and self-reported passage understanding extent ($F(2,202) = 50.967$, $p < .001$, $\eta^2 = .335$). Self-reported passage understanding extent had the highest means in the easy level (Mean = 3.54, SD = .67), followed by those in the hard level (Mean = 3.26, SD = .69), and the medium level (Mean = 2.93, SD = .69), while self-reported passage difficulty degree had the highest mean in the medium level (Mean = 3.23, SD = .72), followed by that in the hard level (Mean = 2.89, SD = .74), and the easy level (Mean = 2.50, SD = .77). All post-hoc p values for the three comparisons (i.e., easy and hard, easy and medium, hard and medium) were lower than 0.01 and salient after Bonferroni correction.

An intriguing observation is that participants reported the highest perceived difficulty level and the lowest perceived understanding extent when reading the medium-difficulty level passages. This may be due to the fact that the medium-difficulty level passages consisted of humanities topics (i.e., archaeology, history, and literature), while the easy and hard passages were primarily science-related (see the Reading Task Section). Humanities passages likely required participants, especially English as second language learners, to expend more cognitive effort in memorizing unfamiliar facts and historical details[59], which may have contributed to the higher cognitive load. This interpretation can be supported by the participant's comments in the post-experiment interview, such as: *"I'm not studying literature or history, so I'm not very interested in reading and cannot memorize most of the content. The content is hard to memorize. History is hard to understand."* (Participant #30 in the BGM group). Accordingly, we excluded the medium-difficulty level and focused on comparing the easy versus hard text complexity levels when answering the aforementioned research question.

Third, we compared each self-reported measure between the two conditions of the low and high comprehension scores which differentiate the hypothesized germane cognitive load. One significant difference was found in self-reported understanding extent ($U = 841.00$, $p = 0.002$, $r = -- 0.352$) between the two conditions of the low and high comprehension scores. Self-reported passage understanding extent in the high score condition (Mean = 3.44, SD = 0.60) was higher than those in the low score condition (Mean = 3.06, SD = 0.52). However, no significant difference was observed in the self-reported passage difficulty level ($U = 1547.50$, $p = 0.095$, $r = 0.192$) between the two conditions of the low (Mean = 2.98, SD = 0.57) and high comprehension scores (Mean = 2.76, SD = 0.66).

Predicting Conditions of Cognitive Load Types by Multimodal Sensor Data and Learner Characteristics

To answer RQ, we examined whether conditions of cognitive load types (extraneous, intrinsic, and germane load) could be predicted by multimodal sensor data and learner characteristics. Using stepwise logistic regression, we identified the most predictive variables that were retained in the final model. Results of the significant prediction models are shown in Table 3. Multicollinearity tests showed that the predictors in each regression model had a low degree of multicollinearity, with all VIF (Variance Inflation Factor) below 2[60].

Higher Fixation Count (B = 0.002, $p < 0.001$) and lower Mean HR (B = − 0.034, $p < 0.001$) significantly predicted the BGM condition while the silence condition was used as the reference. The goodness of fit of the logistic regression model was assessed using a likelihood ratio test, yielding a significant chi-square statistic ($X^2(2) = 50.516$, $p < 0.001$), indicating that the model provides a significantly better fit to the data than a model without the predictors. In addition, the model's R^2 value is 0.082, suggesting that the model explains 8% of the variation in the extraneous load condition.

As noted above, in the subsequent analyses predicting intrinsic load, we excluded the medium-difficulty level and focused on comparisons between the easy- and hard-difficulty levels. Larger Mean Saccade Amplitude (B = 0.454, $p < 0.001$) and higher Fixation Count (B = 0.002, $p < 0.001$) significantly predicted the *hard* category when compared to the reference *easy* category. A likelihood ratio test showed that the regression model with

these predictors fit the data significantly better ($X^2(2) = 29.252$, $p < 0.001$) than the null model, and explained 7% of the variation in the *easy* and *hard* intrinsic load ($R^2 = 0.072$).

Higher Fixation Count (B = 0.001, $p = 0.002$) and greater HRV RMSSD (B = 0.008, $p = 0.033$) significantly predicted high comprehension accuracy, when low comprehension accuracy was used as the reference. A likelihood ratio test indicated that the regression model with these predictors provided a significantly better fit ($X^2(2) = 12.678$, $p = 0.002$) when contrasted with the null model, which explained 2% of the variation in the germane load conditions ($R^2 = 0.021$).

However, the tested learner characteristics (English proficiency, WMC) were not significant predictors for determining the conditions of cognitive load types.

Discussion

Multimodal approach to measure extraneous, intrinsic, germane load

First, we observed that extraneous load type (i.e., the BGM condition with the silence condition as reference) could be significantly predicted by larger eye fixation count and lower mean heart rate (Table 3). The positive relationship between the extraneous load and fixation count can be attributed by the additional cognitive load introduced from listening to their self-selected preferred BGM while reading[10,26,42]. The negative relationship between the extraneous load and mean heart rate can result from the relaxing and mood-lifting effect of participants' preferred BGM since heart rate can reflect participants' anxiety and stress levels[61]. When they experienced a reduction in the stress and anxiety levels, a corresponding decrease in their heart rate became evident[61].

Second, we found that intrinsic load type (i.e., higher text complexity levels while the easy level was used as a reference) was predicted by larger fixation count and mean saccade amplitude (Table 3). As texts with elevated complexity levels typically feature intricate vocabulary and complex sentence structures[36], reading more difficult passages required participants to exert higher cognitive demands in decoding and comprehending the texts[26]. The increased mental workload during reading could be reflected by more frequent fixations[42] and larger saccade amplitudes[62]. Our finding particially contradicts with results in previous reading literature where high cognitive demand was associated with shorter saccade amplitudes[27]. Greater saccade length can indicate that a more extensive region of the stimulus was processed by participants, while smaller saccade amplitude can suggest a more local search[62]. To understand the text with higher complexity, participants in this study might shift their gazes farther for more extensive searches for contextual clues or integrating information, resulting in larger saccade amplitudes[62].

Third, we observed that germane load type (i.e., higher comprehension performance while the lower performance as being the reference) was predicted by larger fixation count and HRV RMSSD (Table 3). As higher fixation count can suggest greater cognitive workload[26,27], one possible explanation for the positive correlation between fixation count and comprehension accuracy is that participants who had higher fixation count might exert more mental effort in reading, such as fixating on words more carefully to ensure full comprehension. In addition, previous research has discovered that HRV measures can indicate cardiac vagal tone, and high vagal tone are associated with fluent cognitive processing in pre-frontal brain areas[63]. Consequently, it can be implied from the positive relationship between HRV RMSSD and germane load that fluent cognitive processing in pre-frontal brain regions might connect to improved comprehension accuracy.

Finally, no significant relationships were found between the tested learner characteristics and hypothesized subtypes of cognitive load, which appears to be incongruent with prior literature[1,19]. One potential reason is that we used stepwise logistic regression models, which retained the most powerful predictors. That is, the learner characteristics (WMC and English proficiency) may have been less predictive than the measures that remained in the models. Another possibility is the relative homogeneity of our sample: all participants were university students with relatively high English proficiency (LexTALE scores: mean = 0.7005, SD = 0.1235, max = 0.9750) and working memory capacity (Two-back scores: mean = 0.8227, SD = 0.1185, max = 1.00), which may have limited the variability needed to observe significant effects[42].

Implications

This paper followed the experiment design of prior studies on types of cognition load and extended it from multimedia learning[1,4] to a new context by examining ESL learners' English passage reading with or without their preferred BGM, which is an important adaptation and contribution to the literature.

On the theoretical side, we showcase that diverse biological metrics such as EM and HR/HRV, which have been previously shown to measure generic cognitive load, have relationships with Sweller's hypothesized three types of cognitive load[6]. The identification of these associations between various proposed measures and the cognitive load types lend support to the validity of the triarchic structure of cognitive load.

Methodologically, previous studies have shown that there might not be a single gold standard behavioural or physiological measure for cognitive load[31]. Thanks to recommendations from MmLA research[2], we took advantage of multiple modalities of data sources that enable researchers to assess cognitive load or understand different types of cognitive load in a more comprehensive manner. In addition, to predict different types of loads with non-intrusive measures, such as EM, HR, HRV, that do not need to halt the ongoing learning process for measurements, has the potential to help researchers better understand the cognitive processes of learners[1,31].

Practically, as EM and peripheral physiological data offer a continuous record of learning processes in real time, this study demonstrates the possibility of incorporating them into computer assisted instruction systems to deliver more adaptive instructions[64]. The computer-assisted instruction systems can empower teachers to provide more timely and appropriate support when multimodal measures detect an increase in learners' cognitive loads, potentially enhance the quality and efficiency of teaching and learning[1].

Limitations and future work

This study was limited to a reading task in which students could autonomously regulate their learning processes and read for an unlimited amount of time with their self-selected preferred BGM. The results may not be generalizable to other kinds of learning tasks (e.g., viewing multimedia lessons). Future studies may further investigate the effects of cognitive load types on other kinds of learning tasks. Second, the self-selected BGM might introduce variability on the caused cognitive load. Future research can examine how specific BGM characteristics (e.g., genre, tempo, presence of lyrics, languages of lyrics) impacts different types of cognitive load. Third, participants in this study were ESL learners from Hong Kong, and thereby the findings may not apply to native English speakers or learners from other parts of the world. Further investigations may explore if the present findings hold true or vary across different learner populations. In terms of individual characteristics, we only considered the participants' general English proficiency and WMC, while future research may take into account learners' executive functions (e.g., attention, multitasking, and planning) which are reported as related to reading comprehension as well[65]. Fourth, we treated germane load as a categorical variable by applying a mean split to the reading comprehension scores to maintain consistency with practices in the extant literature[1,4] and the treatment of extraneous and intrinsic load in this study. A limitation is that this approach may oversimplify individual differences in germane load, particularly for scores near the mean. Future research can consider modelling germane load as a continuous variable to better capture the full range of variation in the participants' learning performance. Fifth, in this study, English proficiency, WMC, and multimodal sensor signals were all treated as personal factors and included as predictors in the logistic regression models. However, English proficiency and WMC can also be modelled as covariates to control for their potential effects, which is a promising direction for future research. Sixth, although the intrinsic load was designed based on Flesch-Kincaid text readability scores, participants perceived the medium-difficulty passages as the most challenging, likely due to the lower topic familiarity. Future work is recommended to incorporate measures of prior knowledge to better align intended text difficulty.

Conclusions

By combining self-reports with biological indicators, this study aims to explore whether and to what extent these measures can predict the three types of cognitive load. Our results showed several intriguing findings: (1) EM metrics were predictive of the three hypothesized cognitive load types; (2) HR/HRV metrics could predict the hypothesized extraneous and germane load. Overall, this study highlights the potential of using automated sensors to investigate cognitive load types in a less intrusive way. This approach has the potential to aid learners by seamlessly integrating into computer-assisted instruction systems or BGM retrieval systems, allowing for dynamical adjustment of the learning environment when learners are facing cognitive overload.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 18 July 2024; Accepted: 19 May 2025
Published online: 23 September 2025

References

1. Zu, T., Hutson, J., Loschky, L. C. & Rebello, N. S. Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *J. Educ. Psychol.* **112**(7), 1338 (2020).
2. Liu, Z., Ren, Y., Kong, X. & Liu, S. Learning analytics based on wearable devices: A systematic literature review from 2011 to 2021. *J. Educ. Comput. Res.* **60**(6), 1514–1557 (2022).
3. Korbach, A., Brünken, R. & Park, B. Differentiating different types of cognitive load: A comparison of different measures. *Educ. Psychol. Rev.* **30**(2), 503–529 (2018).
4. DeLeeuw, K. E. & Mayer, R. E. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *J. Educ. Psychol.* **100**(1), 223 (2008).
5. Sweller, J., Van Merriënboer, J. & Paas, F. Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **31**(2), 261–292 (2019).
6. Sweller, J. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* **22**(2), 123–138 (2010).
7. Paas, F., Tuovinen, J. E., Tabbers, H. & Van Gerven, P. W. Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* **38**(1), 63–71 (2003).
8. Ba, S. & Hu, X. Measuring emotions in education using wearable devices: A systematic review. *Comput. Educ.* **200**, 104797 (2023).
9. de la Mora Velasco, E. & Hirumi, A. The effects of background music on learning: A systematic review of literature to guide future research and practice. *Educ. Technol. Res. Dev.* **68**, 2817–2837 (2020).
10. Que, Y., Zheng, Y., Hsiao, J. H. & Hu, X. Studying the effect of self-selected background music on reading task with eye movements. *Sci. Rep.* **13**(1), 1704 (2023).
11. Du, M., Jiang, J., Li, Z., Man, D. & Jiang, C. The effects of background music on neural responses during reading comprehension. *Sci. Rep.* **10**(1), 18651 (2020).
12. Kiss, L. & Linnell, K. J. The role of mood and arousal in the effect of background music on attentional state and performance during a sustained attention task. *Sci. Rep.* **14**(1), 9485 (2024).
13. Kenz, I. & Hugge, S. Irrelevant speech and indoor lighting: Effects of cognitive performance and self-reported affect. *Appl. Cogn. Psychol.* **15**, 709–718 (2002).
14. Clark, R. C. & Mayer, R. E. *E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning* (Wiley, 2016).
15. Baddeley, A. & Hitch, G. Recent developments in working memory. *Psychol. Learn. Motiv.* **8**, 234–238 (1974).
16. Mayer, R. E. *The Cambridge Handbook of Multimedia Learning* (Cambridge University Press, 2005).
17. Van Gog, T. & Ayres, P. Editorial: State of the art research into cognitive load theory. *Comput. Hum. Behav.* **25**(2), 253–257 (2009).
18. Ayres, P. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learn. Instr.* **16**, 389–400 (2006).

19. Unsworth, N. & Engle, R. W. The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **114**(1), 104 (2007).
20. Gottardo, A. & Mueller, J. Are first-and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *J. Educ. Psychol.* **101**(2), 330 (2009).
21. Proctor, C. P., Carlo, M., August, D. & Snow, C. Native Spanish-speaking children reading in English: Toward a model of comprehension. *J. Educ. Psychol.* **97**(2), 246 (2005).
22. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* Vol. 52, 139–183. (North-Holland, 1988).
23. Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L. & van Merriënboer, J. J. Measuring physician cognitive load: Validity evidence for a physiologic and a psychometric tool. *Adv. Health Sci. Educ.* **22**(4), 951–968 (2017).
24. Paas, F. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *J. Educ. Psychol.* **84**, 429–434 (1992).
25. Van Gog, T., Kirschner, F., Kestner, L. & Paas, F. Timing and frequency of mental effort measurement: Evidence in favor of repeated measures. *Appl. Cogn. Psychol.* **26**, 833–839 (2012).
26. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124**(3), 372 (1998).
27. Dirix, N., Vander Beken, H., De Bruyne, E., Brysbaert, M. & Duyck, W. Reading text when studying in a second language: An eye-tracking study. *Read. Res. Q.* **55**(3), 371–397 (2020).
28. Greco, A., Valenza, G., Lanata, A., Scilingo, E. P. & Citi, L. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng.* **63**(4), 797–804 (2015).
29. Boucsein, W. *Electrodermal Activity* (Springer, 2012).
30. Shi, Y., Ruiz, N., Taib, R., Choi, E. & Chen, F. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* 2651–2656 (2007).
31. Johannessen, E. et al. Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. *Comput. Hum. Behav.* **111**, 106393 (2020).
32. Daley, S. G., Willett, J. B. & Fischer, K. W. Emotional responses during reading: Physiological responses predict real-time reading comprehension. *J. Educ. Psychol.* **106**(1), 132 (2014).
33. Grassmann, M., Vlemincx, E., von Leupoldt, A. & Van den Bergh, O. Individual differences in cardiorespiratory measures of mental workload: An investigation of negative affectivity and cognitive avoidant coping in pilot candidates. *Appl. Ergon.* **59**, 274–282 (2017).
34. Scrimin, S. et al. Dynamic psychophysiological correlates of a learning from text episode in relation to reading goals. *Learn. Instr.* **54**, 1–10 (2018).
35. Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**(2), 175–191 (2007).
36. Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch* (1975).
37. D'Mello, S. K. Gaze-based attention-aware cyberlearning technologies. In *Mind, Brain and Technology* 87–105. (Springer, 2019).
38. Lemhöfer, K. & Broersma, M. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behav. Res. Methods* **44**(2), 325–343 (2012).
39. Lau, E. Y. Y., Eskes, G. A., Morrison, D. L., Rajda, M. & Spurr, K. F. Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. *J. Int. Neuropsychol. Soc.* **16**(6), 1077–1088 (2010).
40. Jainta, S., Blythe, H. I., Nikolova, M., Jones, M. O. & Liversedge, S. P. A comparative analysis of vertical and horizontal fixation disparity in sentence reading. *Vision Res.* **110**, 118–127 (2015).
41. Reichle, E. D., Pollatsek, A., Fisher, D. L. & Rayner, K. Toward a model of eye movement control in reading. *Psychol. Rev.* **105**(1), 125 (1998).
42. Johansson, R., Holmqvist, K., Mossberg, F. & Lindgren, M. Eye movements and reading comprehension while listening to preferred and non-preferred study music. *Psychol. Music* **40**(3), 339–356 (2012).
43. Winter, M., Pryss, R., Probst, T. & Reichert, M. Towards the applicability of measuring the electrodermal activity in the context of process model comprehension: Feasibility study. *Sensors* **20**(16), 4561 (2020).
44. Zhang, L. et al. Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE Trans. Affect. Comput.* **8**(2), 176–189 (2017).
45. Malik, M. et al. Heart rate variability. *Circulation* **93**, 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043 (1996).
46. Shaffer, F. & Ginsberg, J. P. *An Overview of Heart Rate Variability Metrics and Norms* 258 (Public Health, 2017).
47. Makowski, D. et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **6**, 1–8 (2021).
48. Levy, H. E. & Rubinsten, O. Numbers (but not words) make math anxious individuals sweat: Physiological evidence. *Biol. Psychol.* **165**, 108187 (2021).
49. Champseix, R., Ribiere, L. & Le Couedic, C. A Python Package for heart rate variability analysis and signal preprocessing. *J. Open Res. Softw.* **9**(1), 28 (2021).
50. Skervin, T. K. et al. The next step in optimising the stair horizontal-vertical illusion: Does a perception-action link exist in older adults?. *Exp. Gerontol.* **149**, 111309 (2021).
51. Lamelas, M. T., Marinoni, O., Hoppe, A. & De La Riva, J. Doline probability map using logistic regression and GIS technology in the central Ebro Basin (Spain). *Environ. Geol.* **54**, 963–977 (2008).
52. Li, H. Y., He, Y. S. & Li, N. N. The effect of background music on high school students' reading comprehension. *Adv. Psychol.* **2**(4), 206–213 (2012).
53. Thompson, W. F., Schellenberg, E. G. & Letnic, A. K. Fast and loud background music disrupts reading comprehension. *Psychol. Music* **40**(6), 700–708 (2012).
54. Vasilev, M. R., Kirkby, J. A. & Angele, B. Auditory distraction during reading: A Bayesian metaanalysis of a continuing controversy. *Perspect. Psychol. Sci.* **13**(5), 567–597 (2018).
55. Quan, Y. & Kuo, Y. L. The effects of Chinese and English background music on Chinese reading comprehension. *Psychol. Music* **51**(2), 655–663 (2023).
56. George, D., Stickle, K., Rachid, F. & Wopnford, A. The association between types of music enjoyed and cognitive, behavioral, and personality factors of those who listen. *Psychomusicol. J. Res. Music Cognit.* **19**(2), 32 (2007).
57. McFee, B. et al. Librosa: Audio and music signal analysis in python. *SciPy* **2015**, 18–24 (2015).
58. LeBlanc, A., Colman, J., McCrary, J., Sherrill, C. & Malin, S. Tempo preferences of different age music listeners. *J. Res. Music Educ.* **36**(3), 156–168 (1988).
59. Grafstein, A. A discipline-based approach to information literacy. *J. Acad. Librariansh* **28**(4), 197–204 (2002).
60. Kalnins, A. & Praitis Hill, K. The VIF score. What is it good for? Absolutely nothing. *Organ. Res. Methods* **28**(1), 58–75 (2025).
61. Ng, M. Y. et al. Randomized controlled trial of relaxation music to reduce heart rate in patients undergoing cardiac CT. *Eur. Radiol.* **26**, 3635–3642 (2016).
62. Sargezeh, B. A., Tavakoli, N. & Daliri, M. R. Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiol. Behav.* **206**, 43–50 (2019).

63. Thissen, B. A. et al. At the heart of optimal reading experiences: Cardiovascular activity and flow experiences in fiction reading. *Read. Res. Q.* **57**(3), 831–845 (2022).
64. Johnson, E. P., Perry, J. & Shamir, H. Variability in reading ability gains as a function of computer-assisted instruction method of presentation. *Comput. Educ.* **55**(1), 209–217 (2010).
65. Follmer, D. J. Executive function and reading comprehension: A meta-analytic review. *Educ. Psychol.* **53**(1), 42–60 (2018).

Acknowledgements

This research is supported by National Natural Science Foundation of China (No. 61703357) and the Research Grants Council of the Hong Kong S.A.R., China (No. HKU 17607018 and No. HKU 17608621). We thank the contributions of participants to the user experiment.

Author contributions

YQ: Data curation, Formal analysis; Investigation; Software; Visualization; Writing—original draft; Writing—review & editing, YZ: Investigation; Software; Writing—review & editing, JHH: Methodology, Resources; Supervision; Writing—review & editing, XH: Conceptualization, Methodology, Supervision, Validation; Writing—review & editing; Funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03052-1.

Correspondence and requests for materials should be addressed to X.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.