Received: 10 December 2020 | Revised: 26 April 2021 | Accepted: 13 June 2021

DOI: 10.1111/jcal.12590

ARTICLE

Journal of Computer Assisted Learning WILEY

Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity

Katerina Mangaroska[1] | Kshitij Sharma[1] | Dragan Gašević[2] | Michail Giannakos[1]

[1]Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

[2]Faculty of Information Technologies, Monash University, Melbourne, Australia

Correspondence
Katerina Mangaroska, Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway.
Email: katerina.mangaroska@ntnu.no

Funding information
Research Council of Norway, Grant/Award Number: FUTURE LEARNING (255129/H20)

Abstract

Background: Problem-solving is a multidimensional and dynamic process that requires and interlinks cognitive, metacognitive, and affective dimensions of learning. However, current approaches practiced in computing education research (CER) are not sufficient to capture information beyond the basic programming process data (i.e., IDE-log data). Therefore, how cognition and affect intertwine and unfold over time in programming problem-solving activities are rarely investigated.

Objectives: In this study, we examined how the theory-informed measures from multimodal data that we have selected as proxies for cognitive and affective dimensions of learning, are associated with student performance, and in comparison, to prior-knowledge.

Methods: A high-frequency temporal data was collected with a camera, an electroencephalogram, and an eye-tracker from 40 computer science students (bachelor and master studies) in the context of a code-debugging activity. To study the cognitive processes associated with learning we focused on cognitive load theory (CLT) and the human information processing model. In addition, we complemented CLT with the model of affective dynamics in learning to avoid the machine reductionism perspective.

Results: Our findings demonstrated that attention, convergent thinking, and frustration were positively correlated with students' successful code-debugging (i.e., performance), and frequently manifested by high performing participants. Cognitive load, memory load, and boredom were negatively correlated with students' performance, and typically manifested by low performing participants.

Implications: Extending the context of analysis in reference to student cognitive processes and affective states, affords educators not just to identify lower performers, but also to understand the potential reasons behind their performance, making our method an important contribution in the confluence of CER and the learning technology communities. In addition, the insights extracted from our analyses allow us to

discuss potential avenues for improving learning design and the design of interactive learning systems to support the development of students' problem-solving skills.

KEYWORDS
code-debugging task, cognitive-affective states, higher education, multimodal data, multimodal learning analytics

1 | INTRODUCTION

In the last 10 years, higher education has witnessed a substantial increase in the number of learning technologies as a support to the more traditional classroom environments. Contemporary learning technologies afford novel ways for students to learn and instructors to teach, anywhere (across digital and physical settings) and at any time (Clark & Mayer, 2016). As students are free and flexible to choose how they will use learning technologies (e.g., synchronous or asynchronous e-learning), learning in technology-rich settings requires students to apply a diverse set of skills and self-directed learning strategies to successfully assimilate a learning content (Santhanam et al., 2008). However, the different level of skills development, and the various cultural and pragmatic constraints, can cause some students to experience various cognitive challenges (Chew & Cerbin, 2021) and feelings of frustration, boredom, or confusion with the learning content and the tasks, a behaviour that cannot be easily noticed by the instructors in digital settings. Moreover, even if the students care to communicate the challenges and the obstacles they face when learning with technology, the communication is often not in real time (e.g., in a form of a submission, an email). One the one hand, the lag in communication makes it difficult for the instructors to understand the moment when for example, confusion was triggered, for how long, and how frequent during the learning activity. On the other hand, learning technologies rarely have appropriate interventions or feedback mechanisms for the cognitive and affective struggles students face and experience during learning activities. Thus, the benefits from learning technologies in supporting learning and instruction, depend on the extent to which they are compatible with the human affective and cognitive learning processes (Clark & Mayer, 2016). Therefore, increasing our understanding how students' cognition, affect, and behaviour intertwine and span throughout the learning activities, can provide us with valuable insights that can guide the learning design and the development of novel learning technologies.

Whilst important achievements have been obtained in the last decades by mining clickstreams and keystrokes collected through online learning activities (Li et al., 2016; Mousavinasab et al., 2018), learning is ultimately a complex, multimodal process that involves linguistic, gestural, visual, and physical interaction of students with educational systems, learning artefacts, learning spaces, peers, and educators (Kress, 2001; Oviatt et al., 2017; Ritella & Hakkarainen, 2012). Thus, the combination of multiple modalities (e.g., gazing, typing, gesturing) that students employ when learning and communicating, generates rich, objective, and relevant data, comprising of measures that can be assigned as proxies for cognitive and affective dimensions of learning, in the context of problem solving. Problem solving is defined as a 'cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver' (Mayer & Wittrock, 1996, p. 47). The ability to solve complex problems is affected by many factors (e.g., prior-knowledge, type of learning activity, students' set of skills), and requires and interlinks cognitive, metacognitive, and affective dimensions of learning (Mayer, 1998; Spering et al., 2005). However, how these dimensions intertwine and unfold over time in a problem-solving activity are rarely investigated. Thus, to strengthen our understanding how cognition and affect co-exist in situ, and impact the performance of students, we have selected and explored theory-informed learning constructs (e.g., expertise, convergent thinking) salient to problem solving.

In our approach, we have focused on measures extracted from multimodal data, as significant proxies for measuring learning-related constructs (e.g., cognitive load, frustration) in relation to performance. The measures were extracted from students' biomarkers[1] collected with three sensors: an eye-tracker, an electroencephalogram (EEG), and a camera. This way we managed to explore the process of problem solving from two dimensions, that is, cognitive and affective, utilizing measures extracted from the electro- physiological activity of the brain, the facial expressions, the typing and the gaze modality. Moreover, due to the high-frequency of the collected temporal data, we were able to investigate the moment-by- moment tracking of clicks, actions, cognition, gaze, and facial expressions, thereby showing the potential to understand problem solving as a process of change over time. Consequently, our work addresses the following research questions:

1. 'To what extent and how measures extracted from multimodal data that act as proxies for cognitive and affective dimensions of learning are associated with student performance?'
2. 'To what extent prior-knowledge (e.g., expertise) is associated with student performance in comparison to the measures extracted from multimodal data?'
3. 'How measures from multimodal data can inform and influence changes in the learning design?'

In sum, the contribution of the paper is three-fold: *Conceptual* – to advance the discussion on expanding the context and impact of learning analytics research by posing new techniques (i.e., multimodal learning analytics) that can inform and influence changes in the learning design; *Operational* – to deconstruct and investigate a problem-solving learning activity at a fine level of details by utilizing multimodal learning analytics; and *Empirical* – to validate the benefits

of augmenting (i.e., enriching) programming process data (i.e., IDE-log data) with sensor data (e.g., gaze data, facial expressions, EEG data) in the context of learning design.

2 | BACKGROUND

2.1 | Multimodal data and learning

Multimodal data provide information about learners' behavioural (e.g., non-verbal behavioural cues expressed through visual or kinesthetic channels), physiological (e.g., heart rate variability), and mental processes (e.g., cognitive load) that occur during learning activities, and are impossible to be observed and captured with the human eye (Oviatt et al., 2018). Multimodal data can be collected in non-invasive ways using affordable sensor technologies (e.g., eye trackers, wrist-mounted devices, kinetic sensors, electroencephalograms) that monitor variations in different modalities (e.g., speaking, gesturing, gazing, typing) (Lazar et al., 2017). Some applications of sensor technologies in education include: fine-grained analyses of collaborative learning (Malmberg et al., 2019; Martinez-Maldonado et al., 2019), development of real-time feedback mechanisms (Ochoa et al., 2018; Martinez-Maldonado, Echeverria, Schulte, et al., 2020), investigation of self-regulated learning (Azevedo & Gašević, 2019), capturing and studying learning phenomena in classrooms (Chan et al., 2020; Donnelly et al., 2016; Martinez-Maldonado, Mangaroska, Schulte, et al., 2020), teachers' opportunities for reflective practices in relation to data generated from their biomarkers (Prieto et al., 2018), students' emotions in e-learning (Shen et al., 2009; D'Mello et al., 2014) and intelligent tutoring systems (Taub & Azevedo, 2019; Mills et al., 2019).

It has been proposed that multimodal learning analytics (MMLA) has the potential to enable development of models that account for the complexity of the learning process with the purpose of providing real-time feedback (Ochoa et al., 2018), relevant and timely interventions (Blikstein, 2013; Blikstein & Worsley, 2016; Drachsler & Schneider, 2018), and creation of multimodal interfaces (Echeverria et al., 2019; Martinez-Maldonado, Echeverria, Fernandez Nieto, & Buckingham Shum, 2020), to name a few. In fact, there are MMLA studies that have been focusing on modelling student gaze to identify group synchrony as a proxy of collaboration effectiveness (Schneider, 2020); capturing physiological cues to investigate group regulation strategies (Noroozi et al., 2019) and individual achievement (Pijeira-Díaz et al., 2018); utilizing computer vision systems to identify incorrect postures in healthcare training (Di Mitri, 2019); creating hand tracking algorithms to predict group work quality (Spikol et al., 2018); and using positioning trackers to identify teaching strategies in physical classrooms (Martinez-Maldonado, Mangaroska, Schulte, et al., 2020; Martinez-Maldonado, Echeverria, Schulte, et al., 2020).

The findings from all these studies demonstrate that MMLA can derive a more comprehensive view of learners' behaviours, actions, cognitive and affective states, as well as model meaningful learning constructs from commonly intertwined data-markers (e.g., heart rate, gaze, cognitive workload, stress level, and arousal) salient to learning. However, most of these studies portray phenomena without sufficient use of theory (Gašević et al., 2015), and without analysing the learning activity at the level of details we have considered in this study. To that end, the work in this paper focuses on: (1) measures informed by literature and grounded in theory, to minimize the risk of establishing weak concepts or missing to identify other important patterns in the data; and (2) exploring problem solving as a dynamic process and not an outcome, by tracking clicks, actions, cognition, gaze, and facial expressions from moment-to-moment.

2.2 | Problem solving and learning

Problem solving is a dynamic process that unfolds in different phases over time. A recent fMRI study has established the existence of three learning phases, namely encoding, solving, and responding (Tenison et al., 2016). Problem solving is considered to be the bridge between learning and performance (Anderson, 1993), because it entails many behavioural and cognitive multi-step activities (governed by metacognitive awareness and emotions), that convert what is learned into behaviour and towards performance (e.g., goal attainment; Dörner & Funke, 2017).

When solving a problem, learners are required to apply higher-order cognitive skills, such as divergent and convergent thinking (Johnson, 1997). Divergent thinking is used to generate ideas to a particular problem; however, without *convergent thinking*, learners cannot select and organize the information to converge on a correct solution (Csikszentmihalyi, 1996; Chang et al., 2016). To do so, the learner's mind requires *sustained attention* (i.e., concentration) and mental capacity to process new information considering the working memory constraints (Wang et al., 2013). To avoid 'overloading' their cognitive system, learners direct their attention to *specific parts* [e.g., *area of interest (AOI) du- ration*] to select relevant information, guided by their cognitive strategies and metacognitive awareness. However, if there is a discrepancy between the *memory load*[2] and the *mental effort*,[3] a learner can experience increase in the *cognitive load*[4] (Sweller et al., 2019). In other words, cognitive load is not affected only by the characteristics of the task (i.e., the learning design and the interface design), but also by the characteristics of the subject performing the task, and the interaction between the two (Paas & Van Merriënboer, 1994).

To describe the cognitive processes associated with learning (Paas & Van Merriënboer, 1994; De Jong, 2010) and to model the cognitive aspects of human behaviour (Hollender et al., 2010), researchers in the field of instruction and learning often look into *cognitive load theory* (CLT). On the one hand, furthering our understanding about problem solving in digital settings through exploration and measurement of cognitive processes (Anderson, 2013; Razoumnikova, 2000), can reveal what measures from various modalities can be mapped back to pertinent learning constructs, thereby establishing measures grounded in theory. Such mapping of measures and constructs can influence and enable improvements in the learning design, that can bring on the development of learning activities to extend the human cognitive capacities and learning abilities during

problem solving (Paas & Van Merriënboer, 1994; Mayer & Wittrock, 1996). On the other hand, such measures can be set for estimating learners' cognitive load and mental load in computer-mediated learning activities, which can have a practical value for future design of human-centred adaptive and interactive learning systems (Chen & Epps, 2014; Haapalainen et al., 2010).

Although the constructivist view of learning focuses on cognitive changes within learners, problem solving requires and interlinks cognitive, metacognitive, and emotional dimensions of learning (Mayer & Wittrock, 1996; Jackson et al., 1996; Spering et al., 2005; Dörner & Funke, 2017). Hence, we have extended our approach for selection of learning constructs salient to problem solving, to *the model of affective dynamics in learning* proposed by D'Mello & Graesser (2012). This model emphasizes the role of *cognitive disequilibrium* in learning and problem solving. According to this model, when learners solve complex problems and face an error or are uncertain what to do next, they enter in a state of cognitive imbalance which is accompanied by the affective state of *confusion* (D'Mello & Graesser, 2012). The state of confusion triggers reasoning and reflection, so that learners can restore the state of *cognitive equilibrium*. If the learners cannot resolve the issue, they experience *frustration*, which if it is persistent, it can easily transition into *boredom*, a point when learners disengage from the learning process (D'Mello & Graesser, 2012). On the other hand, a learner in a state of flow and engagement, often exhibits high degree of satisfaction (i.e., *delight*), a positive affective state that has complementary effect on broadening the scope of attention (Fredrickson & Branigan, 2005).

Building on CLT (De Jong, 2010) and the model of affective dynamics in complex learning activities (D'Mello & Graesser, 2012), we have selected *theory-informed learning constructs* (summarized in Table 2) to further our understanding how cognition and affect co-exist, and impact the performance of students, thereby extending our knowledge of problem solving in digital settings, and informing the learning design and the design of computer-mediated learning environments.

2.3 | Psychophysiology and problem solving

Psychophysiology is the study concerned with 'the measurement of physiological responses as they relate to behaviour (e.g., problem solving, information processing)' (Andreassi, 2010, p. 44). Psycho- physiological measures carry many challenges with respect to privacy, invasiveness, sensitivity, interpretability, and generalizability (Andreassi, 2010); thus, their broad application in learning and teaching is yet to be seen.

One early example of employing psycho-physiological measures to investigate problem solving, is the study by Aula & Surakka (2002) who explored the effect of emotional feedback on human behaviour in a computerized problem-solving math task. They found that positive feedback triggers significantly faster decrease in the pupil diameter (which is a measure linked to cognitive load) than negative or neutral feedback. Such insights can lead towards the design of methods for emotions regulation in humans for e-learning. Yoon & Narayanan (2004) used gaze-related measures to explore trajectories of users' visual attention strategies during problem solving. The insights from this study have practical implications for designing user interfaces that can guide users' visual attention and reduce the cognitive load inherent in mental imagery, by providing additional information that reduces the response time and increases accuracy (i.e., improves users' problem-solving performance). Similar to Yoon & Narayanan's (2004) study, Mangaroska et al. (2018) utilized gaze data to explore visual attention strategies among novices and experts in problem-solving programming activity. Their findings showed that measures from multimodal data can be used to develop tools that can orchestrate basic behaviour regulation (e.g., how a user processes information or interacts with visual information), and as such, guide students to attend the right information at the right time to maximize the understanding of relevant concepts.

Considering the relation between task dependency and gaze patterns, Kaller et al. (2009) conducted a study to gain a better understanding of visuospatial problem solving. The results demonstrated task- dependent eye-movement patterns, supporting a sequential model of problem solving as internalization, planning, execution, and verification. More recently, Tenison et al. (2016) conducted an fMRI study in which they established the existence of three qualitatively distinct learning phases during problem solving utilizing brain-related measures. The authors advanced the understanding of skill acquisition when solving a novel complex math problem with repetitions that eventually can help with pre- and post-training study designs.

All these studies advance the current body of knowledge in the context of problem solving, by studying problem solving as a dynamic process of change and by deriving insights obtained with sensor data, as data highly representative for the human affective and cognitive learning processes. This brings the community closer to understand the evolution and the sequence of different phases of problem solving, as well as the co-existence of cognitive, affective, metacognitive, and motivational dimensions of learning.

3 | METHODOLOGY

This section presents the design of the study, the methods used to collect and process the data, and the employed analysis approach to address the research questions.

3.1 | Research design

The research design of our study is a single-group time series design (Ross & Morrison, 2004) involving repeated measurement of a group with the experimental treatment induced. Our study consists of a debugging as the treatment, continuous measures (via behavioural log data and the multimodal data shown in Table 2) as predictor variables, and the performance captured thought students' progress with the task as the dependent variable. We decided to use single-group time series because the collected observations were gathered through repeated measurements over time (i.e., the measures from multimodal data presented in Table 2 were tracked, monitored, and aggregated over time). The time series design is suitable for detecting unstable

and temporal behaviour patterns. Moreover, the effect of our experimental treatment is likely to be more apparent in a repeated measurement design due to the unsystematic variance that can be caused by the changes in students' behaviour over time, while keeping the 'noise' to a minimum. In particular, we designed and implemented a code-debugging task to explore learning constructs associated with problem solving, that are informed from theory and literature, and frequently applied indicators in education and problem-solving research. The main task covered debugging a Java class named Person (that manages parent-child relationships), accompanied with five questions, written right after the code, presented as a part of the main method. The code provided to the participants tried, but failed to ensure consistent object relationships (Figure 1).

3.2 | Deconstructing the code-debugging learning activity

To further our understanding about problem solving, we deconstructed the code-debugging process in three main phases: understanding, changing-testing (i.e., finding bugs and testing the code), and fixing. This decomposition corresponds to a recent fMRI study, that established the existence of three qualitatively distinct learning phases during problem solving, namely encoding, solving, and responding (Tenison et al., 2016). For each of the phases we looked into the programming process data, that is, IDE-log data (i.e., behavioural dimension) and the fixation-duration on the defined AOIs (i.e., behavioural dimension). On top of the behavioural log data, we added the measures we extracted from the multimodal data (see Table 2).

Next, we defined two types of behavioural actions: reading (R) and writing (W) episodes. The reading episodes covered actions when students were reading the code, the output, or the assigned questions, while the writing episodes covered actions when students where editing the code to check the output or the questions (e.g., students were commenting the questions or writing notes for themselves). To map these actions back to the three phases of the code-debugging process, we divided the episodes into initial reading (Ri) and writing (Wi) episodes, and later (i.e., subsequent) reading (Rn) and writing (Wn) episodes. The initial episodes were the first 10% time duration for each debugging question (see Figure 2 for computation of the initial and later R-W episodes). Consequently, Ri and Wi were mapped onto the first stage (i.e., understanding), where students were reading throughout the code and the questions (Ri), and were making small edits in the code or the print statements in the questions (i.e., Wi) to check the initial output. Rn and Wn were mapped onto (1) the second stage (i.e., changing-testing) where students were demonstrating the continuous 'loop' of changing the code, evaluating the hypotheses, and testing their written solutions by running the main method; and (2) the third stage (i.e., fixing) where students were changing the code to fix the already located bugs. The research design is shown in Figure 3

3.3 | Participants and procedure

During the spring semester 2019, the experiment was performed at a contrived computer lab at the *Norwegian University of Science and Technology (NTNU)*, with 46 students (8 females and 38 males), age between

FIGURE 1 Graphical representation of the code-debugging task. The consistencies that were absent from the original version of the code. (1) Gender consistency: the mother should be a female and the father should be male. (2) Child-parent consistency: if Jens is the child of Merit, Merit should be the mother of Jens; and vice-versa. (3) The removal of a child-parent relationship from either a parent or a child should also apply to the whole family. (4) Adoption consistency: the child-parent (addition and removal) and the gender consistencies should be maintained in the case of an adoption

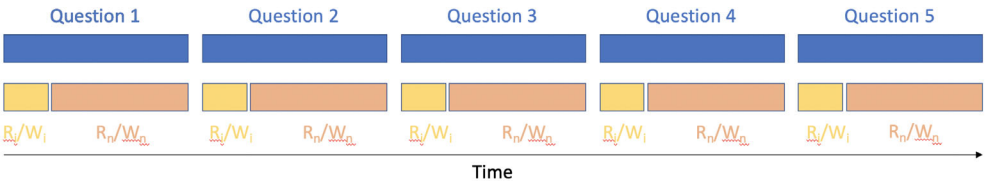

FIGURE 2 The approach used to calculate initial and later R-W episodes

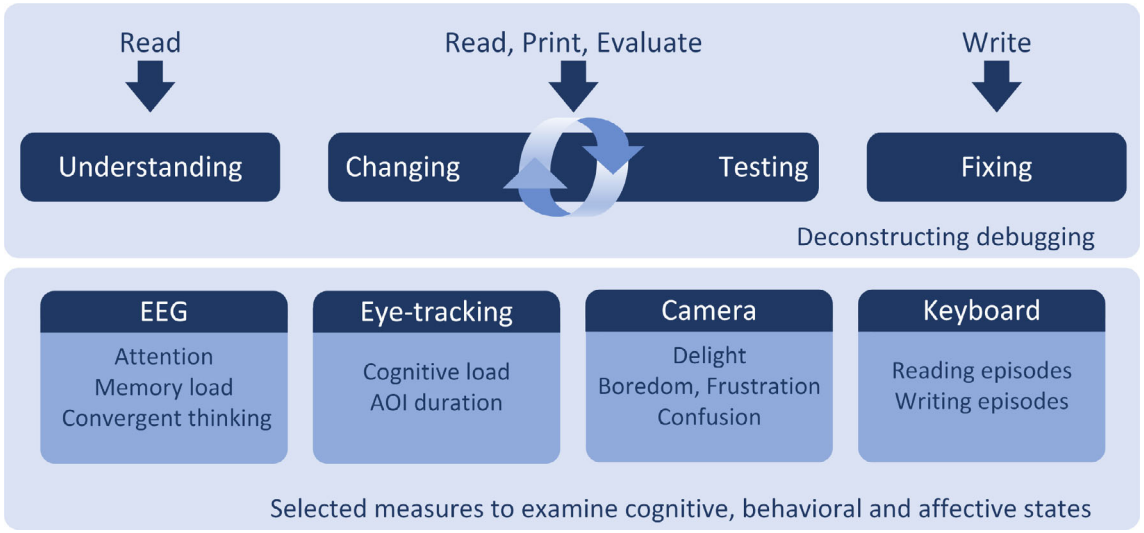

FIGURE 3 Outline of the research design

20 and 25 ($M = 22.1$, $SD = 1.46$). The students were recruited from all study years of the computer science major degree via mailing list. We did not recruit students in their first year because they had not taken yet a course in object-oriented programming (OOP). All recruited students had used Eclipse IDE during their OOP course. The experiment ran for a week – a total of 20 non-repeat sessions, where each session had two students at a time, on two separate computers. The students were instructed not to talk to each other, which was verified through the video data recorded with the cameras. At the end of the experiment, students received a gift voucher equivalent to 30 euros for their participation.

Upon arrival in the lab, the students were briefed about the experiment following the basic ethical principles suggested by the Department of Health (2014). The briefing included the following: (1) the experiment is based on a voluntary participation; thus, students could opt-out at any moment; (2) there is no risk of harm (physical or psychological) from using the sensors; (3) their privacy will be protected and guaranteed; and (4) their individual data will be anonymized and aggregated before any analyses could materialize. Detailed information about the experiment was also provided in the consent form, that the students signed following the briefing. Then, the lead researcher explained the sensors that were used during the experiment, and placed an EEG ENOBIO cap on students' heads. Next, the eye-trackers were calibrated using a 5-point calibration process, while the EEG was calibrated using the off-the-shelf ENOBIO EOG correction software.[5] After the calibration process, the students were asked to finish three small code-debugging assignments (easy, medium, difficult) within 20 min. We considered this as a pre-task test, which was used to decide the students' expertise. Then, the students were given 40 min to solve the main task. The code for the main task contained no syntactic errors, and the students were notified about this fact. The stages are shown in Figure 5 and the whole set-up of the experiment is presented in Figure 4.

3.4 | Data collection

During the learning activity (pre-task and main-task), we collected data from four sources: an EEG device (i.e., brainwave signals), an eye-tracker (i.e., gaze data), a camera (i.e., video data with participants' faces), and programming process data (an IDE-log data). The collected data included activity (e.g., logs), neural (e.g., electrophysiological activity of the brain), and natural communication patterns (i.e., gaze data, facial expressions) (Oviatt et al., 2018). All sensor data were synchronized by having all devices' clocks synchronized with the computers that participants were using. Table 1 shows the dimensions of the physiological data.

The data collection process for each of the data streams is described in the following:

1. *EEG data*: The EEG signals were recorded with a 20-channel ENOBIO device following the international 10–20 system, as shown in Figure 4. The raw EEG signal data were recorded at a 500 Hz using a portable EEG cap, and divided into the following

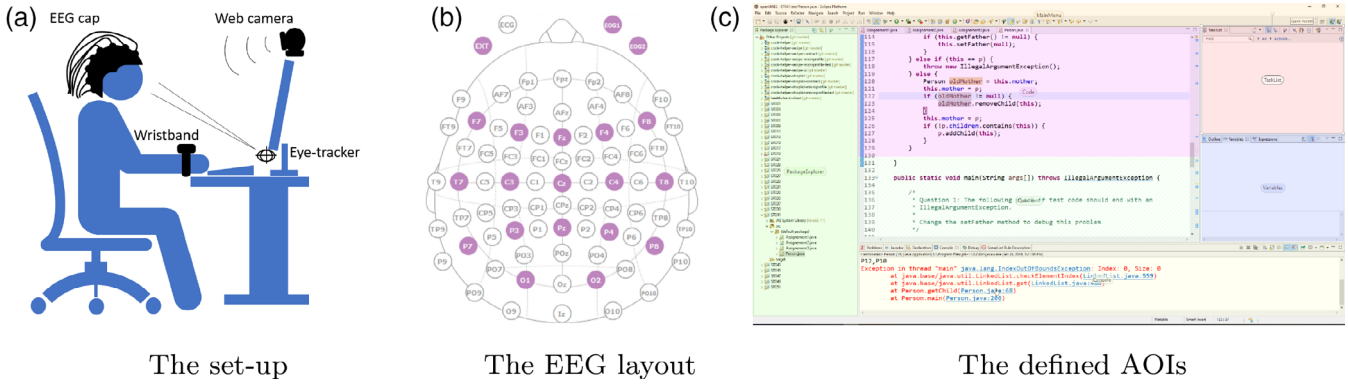

FIGURE 4 Design of the experiment. (a) The set-up, (b) the EEG electrode layout of 20 channels, and (c) the defined AOIs in Eclipse. The standard electrode layout shows the coloured electrodes that are being used in the experiment, and the white electrodes that ENOBIO provides option for. AOI, area of interest; EEG, electroencephalogram

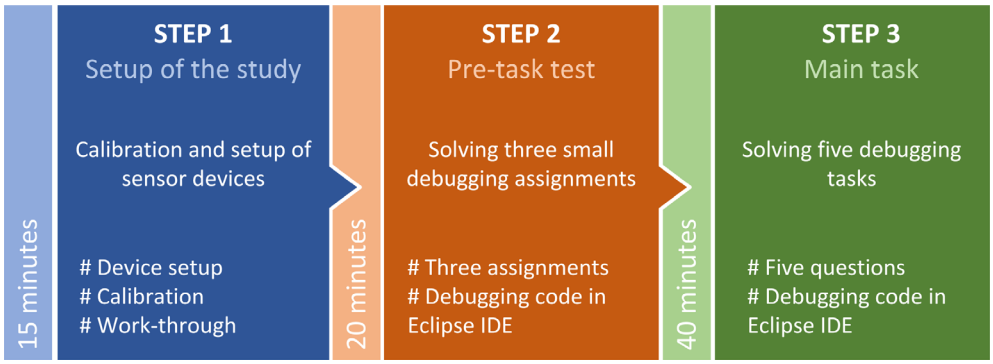

FIGURE 5 The three stages of the experiment

TABLE 1 Dimensions of the physiological data (Andreassi, 2010)

Source	Data	Units	Sampling frequency
EEG	Alpha – α	Hz	8–12 Hz
EEG	Beta – β	Hz	18–30 Hz
EEG	Theta – θ	Hz	4–7 Hz
Eye-tracking	Pupil diameter	mm	120 Hz
Eye-tracking	Fixations	ms	120 Hz
Eye-tracking	Saccades	ms	120 Hz

Abbreviation: EEG, electroencephalogram.

band powers: theta – θ (4–7 Hz), alpha – α (8–12 Hz), and beta – β (18–30 Hz; Haapalainen et al., 2010). The Fz electrode was used as a signal reference electrode, two channels were used for EOG correction, one channel for electric reference, and three Channels Accelerometer with sampling rate at 100 Hz.

2. *Gaze data*: To record students' gaze, we used a Tobii X3-120 eye-tracking device at a 120 Hz sampling rate and using a 5-point calibration. The device is non-invasive and mounted at the bottom of a computer screen. The screen resolution was 1920 x 1080 and the students were 50–70 cm away from screen. All students sat on a non-wheeled chair in front of the computer screen.
3. *Facial expression data*: To capture face expressions from the students, we used LogiTech web camera, pointed straight at the students from the screen, capturing video at 30 frames-per-second (FPS). The web camera focus zoomed at 150% onto the faces of the students. During the tasks, the students exhibited a minimal body and gesture interaction; hence, the video recordings hold high quality data from students' facial expressions. The video resolution was 640 x 480.
4. *Log data*: An Eclipse plug-in, that is, an exercise view (Trætteberg et al., 2016), was used to gather the reading and writing behaviour of the students. This plug-in captures the state of the programme every time a student saves the programme, either by clicking on the 'save' button or by pressing 'CTRL+S'.

3.5 | Data pre-processing

The raw data from the sensors' recordings contain artefacts as a result of (1) the blinks (e.g., gaze data); and (2) the adjustments of the EEG cap, the jaw movements, and the blinks (e.g., EEG data). To prevent distortions in the analysis and to ensure validity, we were required to detect and remove such artefacts. Due to missing data, calibration errors, and temporal mismatch, we removed six participants from the

TABLE 2 Theory-informed learning constructs

Learning constructs	Meaning	Learning dimension	Measures	Data stream
Attention	State of arousal when humans selectively concentrate on a discrete aspect of information.	Cognitive	α band power	EEG
Convergent thinking	Convergent processing of internal attention directed to one correct task solution.	Cognitive	upper β band power	EEG
Memory load	Composite of demands placed on the working memory capacity by the task during memory retention.	Cognitive	θ band power	EEG
Cognitive load	The load that performing a task imposes on the cognitive system of a learner, considering casual and assessment factors.	Cognitive	index of pupilary activity computed as discrete wavelet transform of the pupil diameter	Eye tracking
AOI duration	Interaction time with the IDE.	Behavioural	proportion of time looking at the screen.	Eye tracking
Delight	High degree of satisfaction.	Affective	AU4, AU7, AU12, AU25, AU26	Face
Frustration	Dissatisfaction or annoyance from being stuck.	Affective	AU12, AU43	Face
Boredom	Being weary or restless through lack of interest.	Affective	AU4, AU7, AU12	Face
Confusion	Lack of understanding and being unsure how to proceed.	Affective	AU1, AU4, AU7, AU12	Face
Reading episode	Reading lines of code.	Behavioural	>30 s, mean and s.d.	IDE-logs
Writing episode	Editing lines of code using the keyboard.	Behavioural	<30 s, mean and *SD*	IDE-logs

Note: Reference for the borrowed method: attention (Cooper et al., 2006; Klimesch et al., 1998), convergent thinking (Shemyakina & Dan'ko, 2007; Zhou et al., 2019), memory load (Jensen & Tesche, 2002; Grunwald et al., 1999), index of pupilary activity (Duchowski et al., 2018), AOI duration (Holmqvist et al., 2011), expressions from action units (Baltrušaitis et al., 2016), affective states (McDaniel et al., 2007), reading-writing episodes (Sharma et al., 2018). Abbreviation: EEG, electroencephalogram.

data set. For the rest of the participants we have cleaned the data in the following manner:

1. *EEG data*: First, an independent component analysis (ICA) was used to remove the noise from the jaw movements.[6] This was accomplished using separation of the signal into signal and noise, where the noise was set to be coming from the jaw movements of the students. We also applied an EOG filter (in-build function in the ENOBIO software for neural data processing) to remove the noise from the blinks and the eye-brow movements, and an additional filter to remove the noise from the tongue movements. A 60 Hz line filter was also used to remove any noise coming from the interference within the EEG wires.
2. *Gaze data*: Tobii's default algorithm (i.e., in-build function in the Tobii software for gaze data processing) was used to identify fixations and saccades (for details please see Olsen, 2012). A filter (i.e., in-build function in the Tobii software) was used to remove the raw gaze points that were classified as blinks.
3. *Facial expression data*: In most of the frames in the video recordings only one face was visible. However, sometimes the lead researcher appeared in the field of view of the camera. Due to the settings of the experimental space, the researcher could only appear to the right side of the student. Moreover, the algorithm in the OpenFace recognition software gave each face in the frame an ID from left to right. This means that in the frames where both the researcher and the student were present, the student's face ID was always zero. For frames with two faces (as this was the highest number of faces in any frame) the researcher's face that had an ID value of one was systematically removed.

Another important issue with physiological data is the susceptibility of the data to various personal and contextual biases. Examples of these biases include: time of the day, physical health condition of the participants, gender, age, and an overnight sleep quality. All data, except the facial expression data, were normalized using the first 30 s of the data streams, to remove the subjective and contextual bias. Thus, for normalization, every data point was expressed as a proportion of the means of the first 30 s. Further, the pupil dilation was also normalized with the darkest and the brightest screen shots in the whole interaction for each student (Armato et al., 2013). Finally, the time series were normalized using the MinMax normalization.Next, the data were divided into small episodes of up to 30 s each. Then, all measures (shown in Table 2) were computed. Considering the *behavioural dimension* of learning, AOIs were calculated as a proportion of the time students spend looking at the different areas of the screen, and the R-W episodes were computed depending whether they lasted more or less than 30 s. Considering

the *cognitive dimension*, we computed the band power of each frequency band (i.e., α, upper β, and θ) in each time window, while the cognitive load was calculated as the index of the pupilary activity (Duchowski et al., 2018). To get to the band power, first we performed a Fast Fourier Transform (FFT), and then we blocked all the frequencies higher or lower than the bandwidth (using a band pass filter). Then, we converted the remaining signal to time domain by using an inverse FFT, to compute the band power as the root mean square of the amplitudes. At last, considering the *affective dimension*, we computed the action units that corresponded to each of the learning-related constructs (i.e., delight, confusion, frustration, and boredom) (McDaniel et al., 2007). All computed measures (see Table 2) were aggregated based on where the student was looking (i.e., AOI) and whether it was a reading or a writing episode. The features were computed in a temporal manner, and to correspond to the deconstructed code-debugging process we chose to summarize them. This way we kept the short (in time) intricacies and compare them over a longer period of similar behaviour. Each data point corresponds to one student for each of the AOI (reading/writing) episode; thus, we kept the analysis of variance (ANOVA) assumption about the independent sampling.

3.6 | Variables

The variables we have selected to explore are theory-informed and in relation to the relevant body of work (please see Table 2). All variables are continuous and frequently applied in education and problem-solving research (please see Section 2). We have selected the *debugging performance* (from here on performance) to be our dependent variable. Moreover, as expertise is a complex phenomenon that is highly contextualized and develops over time (McCauley et al., 2008), it was necessary to examine if expertise has influence on students' performance and how much explanation power added it to the models. The following are the rest of the variables, which we have selected to be the experimental variables:

Expertise: The expertise of the students was decided from the pre-task which consisted of three small code-debugging tasks. Each task contained three bugs with the same level of difficulty. We expected students to remove all bugs within 20 min. The expertise ranged between 0 and 3, depending on the number of tasks successfully solved by the students. For example, three bugs per assignment needed to be fixed, so that the assignment could be counted as solved. The following are the percentages of the students that solved none, one, two, or three tasks: 0–40%; 1–25%; 2–27.5%; 3–7.5%.

3.6.1 | Debugging performance

To finish the main code-debugging task the students were required to solve the five questions in a particular order. Students were given 40 min to complete the main task. At the end of the 40 min, they were asked to stop, and the number of solved questions at that point of time, was taken to be the measure of performance. The performance ranged between 0 and 5, depending on the number of solved questions by the students. The following is the percentage of the students that solved none, one, two, three, four, or five questions: 0–17.5%; 1–2.5%; 2–17.5%; 3–10.0%; 4–32.5%; 5–20%.

3.6.2 | Individual areas of interest

Eclipse IDE was divided into seven functional AOIs (these are the basic panels in the interface of the IDE) shown in Figure 4. For the analysis, we have computed the proportion of time students spent on three AOIs: Code, Output, and Questions, as previous studies have shown these AOIs to be particularly important in code-debugging tasks (Bednarik, 2012; Mangaroska et al., 2018). Moreover, project explorer and toolbar did not include any information important for comprehending and solving a task; DebugView was not used by any of the students, and VariableView was used only by few students.

3.6.3 | Code reading and writing episodes

The reading-writing actions depict the difference between the time when students are editing, versus the time when they are only reading the code. The writing episodes were detected using the activity of the keystrokes. An uninterrupted typing (with breaks smaller to 30 s) segment was annotated as a writing episode, while no typing activity longer than 30 s (i.e., a data-driven threshold), was annotated as a reading episode. We computed mean and standard deviation for these reading-writing actions.

3.6.4 | Theory-informed learning constructs

Table 2 summarizes the constructs that were computed from the multimodal data. The table provides the learning constructs, their meanings, the learning dimension each construct covers, how they have been measured, the data stream employed, and the respective literature source from where the constructs have been adopted. In particular, attention, convergent thinking, and memory load were computed using the different frequency bands of the EEG data. Cognitive load was computed using the eye-tracking data, and the facial expressions (i.e., delight, frustration, boredom and confusion) were computed using the facial video data. All these measurements either capture a specific process during learning and/or problem solving, or have been found to be related to learning/problem-solving performance. For example, convergent thinking is often correlated with performance in tasks that have closed-ended solutions (e.g., debugging, programme comprehension, puzzles) as compared to the open-ended tasks (e.g., creative tasks and brainstorming; Zhang et al., 2020). The upper beta band (18–30 Hz) of EEG increases with the increase in the convergent thinking in various tasks (Zhou et al., 2019; Shemyakina & Dan'ko, 2007). Similar to convergent thinking, attention has been found to be positively correlated with learning performance in several

studies (Chen & Wu, 2015; Sharma et al., 2014). When specifically measured with EEG (Benedek et al., 2014; Cooper et al., 2006) it was found that the alpha band (8–13 Hz) power increases with participants' attention in tasks, thereby, showing that the alpha band power is also related to attention specific errors (Carp & Compton, 2009). On the other hand, memory load and cognitive load have been negatively correlated to the learning performance (Sprague et al., 2014; Wang et al., 2018). Higher cognitive and memory load might trigger disengagement from a task (Boekaerts, 2017; Gordon et al., 2014) and in turn, be detrimental for the performance (Bergdahl et al., 2020).

Finally, the affective dimensions (i.e., delight, boredom, frustration and confusion) have been related to various learning processes (D'Mello & Graesser, 2012). During learning/problem solving, confusion occurs when the groups have to reinforce their pre-existing mental models with new information (Clarebout & Elen, 2001; D'Mello & Graesser, 2012). On the other hand, frustration, during learning sessions, was found to be eminent in online interaction (Capdeferro & Romero, 2012) and in online discussion forums (Chen & Caropreso, 2004). Frustration and confusion were shown to lead to impasses in problem solving (VanLehn et al., 2003). Lastly, when the problem at hand is far too easy or repetitive, boredom was the emotion that was mainly observed in past studies (Panitz, 1999; Baker et al., 2010). Based on a selective meta-analysis with 21 studies (D'Mello, 2013), in this paper, we decided to focus on these expressions because they were found to be most prominent for complex learning activities.

3.7 | Data analysis

To answer RQ1, we have created two linear models (Table 3, model 1 and model 3), with the performance as the dependent variable and three sets of independent variables: (1) the measures from multimodal data (Table 2); (2) the gaze on the different AOIs; and (3) the initial episodes (reading/writing) for the first model, and the later episodes (reading/writing) for the second model. To answer RQ2, along with the first set of models, we have created another set of two models (Table 3, model 2 and model 4), which include a new independent variable, that is, the expertise. To find the relation between the expertise and the performance we have used Pearson correlation. In all the models, the significance of the coefficients was tested using a two-tailed *t*-test.

The models that we have created were based on the segmentation of the eye-tracking (i.e., AOIs) and the IDE-log (i.e., R-W episodes) data streams. The segments were based on two factors. First, whether the learners were reading or writing, and second, which AOIs they were looking at. Hence, we segmented the whole interaction in initial and later (i.e., subsequent) R-W sessions based on the AOI durations, and we have created four linear regression models (including the models with the expertise as an independent variable). For all models, the debugging performance was the dependent variable, and the selected measures shown in Table 2 were the independent variables.

To answer RQ3, we have analysed the most significant multimodal data measures from the above-mentioned models. Considering previous findings outlined in the cognitive and affective research, and the findings from our study, we have proposed learning design guidelines that can aid educators to scaffold the code debugging process and thereby, augment students' debugging performance.

TABLE 3 Details for the different models

Model ID	Independent variables	Predictors
Model 1	Debugging performance	Measures from Table 2, AOI being looked at, Initial episode type (reading/writing)
Model 2	Debugging performance	Measures from Table 2, AOI being looked at, Initial episode type (reading/writing), Expertise
Model 3	Debugging performance	Measures from Table 2, AOI being looked at, Later episode type (reading/writing)
Model 4	Debugging performance	Measures from Table 2, AOI being looked at, Later episode type (reading/writing), Expertise

Abbreviation: AOI, area of interest.

4 | RESULTS

First, we analysed the relation between expertise and performance. Because there was a positive and significant correlation between expertise and performance ($r(40) = 0.56$, $p = 0.0001$), the rest of the analysis focused on the performance as a dependent variable. We also compared the models with and without expertise, as an additional independent variable. Table 4 shows the adjusted *R*-squared values (adjusted for the additional estimation of the parameters) for each of the four models. One can observe that including expertise as an additional independent variable did not add much extra information to explain the variance in the performance, which supports our decision to discard the expertise as an independent variable from the models. Although one might argue that multimodal data models are as good as expertise and thus, use expertise as a distinguishing variable; having a pre-test to measure expertise is not always possible or practiced by the instructors. Therefore, we propose to use the models without expertise and with the features extracted from the multimodal data streams.

Tables 5 and 6 are showing the most significant results from the linear regression, calculated for the two models according to the initial and later R-W episodes and the AOI durations, in relation to performance. As one can notice, when the students were in the *understanding phase* (i.e., *encoding*), once they started encoding, most of the

TABLE 4 Adjusted R^2 for the four models with and without expertise for modelling performance

Adj. R^2	W/O expertise	With expertise
Model 1 and Model 2	0.71	0.74
Model 3 and Model 4	0.74	0.76

TABLE 5 Results from the linear regression based on initial R-W episodes

Initial R-W episodes	AOI	Learning-related construct	Estimate	SE	t-Value	p-Value
		(Intercept)	0.40	0.25	0.53	0.52
Initial reading episode	Code	Cognitive load	−1.01	0.0051	−2.07	0.05
		Attention	1.27	0.0016	2.49	0.01
		Convergent thinking	1.56	0.0002	2.08	0.05
		Boredom	−2.95	0.0388	−2.79	0.01
		Delight	−1.81	0.0303	−2.96	0.01
		Frustration	1.13	0.07	2.13	0.05
		Memory load	−2.68	0.0052	−2.03	0.05
	Question	Attention	3.56	0.0026	3.53	0.001
		Convergent thinking	0.78	0.0062	1.97	0.05
		Boredom	−0.79	0.0271	−2.16	0.05
	Console	Convergent thinking	1.69	0.0042	2.06	0.05
		Delight	−0.81	0.0530	−2.19	0.05
		Frustration	1.17	0.0667	2.37	0.05
Initial writing episode	Code	Cognitive load	−1.22	0.0168	−2.32	0.05
		Memory load	−1.08	0.0012	−2.01	0.05
		Attention	1.17	0.0022	2.17	0.05
		Convergent thinking	0.98	0.0001	2.92	0.01
		Boredom	−1.40	0.0518	−2.70	0.01
		Delight	−0.90	0.0303	−2.97	0.01
		Frustration	1.52	0.0642	2.33	0.05
	Question	Cognitive load	−2.05	0.0079	−2.70	0.01
		Convergent thinking	1.85	0.0066	2.05	0.05
		Boredom	−1.71	0.0919	−2.58	0.01
	Console	Confusion	0.75	0.0158	2.17	0.05
		Convergent thinking	3.65	0.0006	2.93	0.01

cognitive-affective states they displayed were related to performance, except *memory load* and *confusion*. As they progressed towards reading the questions (Ri) and making small edits in the code (Wi), we observed a negative correlation between *memory load* and performance, and a positive correlation between *confusion* and performance when students looked in the console AOI for the outcomes of those small edits in the code. *Convergent thinking* was positively correlated to the performance, and was observed in all AOIs (i.e., code, question, console) in the initial R-W episodes, which we have considered it as a sign that the students were actively engaged in solving the task. *Frustration* was positively correlated to the performance, and demonstrated by the high-performing participants as a sign that they were more actively engaged in solving the code-debugging task than the low-performing participants.

In the later R-W episodes, which include the *changing-testing* (i.e., *solving*) and *fixing* (i.e., *responding*) *phases*, the students continued to demonstrate their active engagement with the code-debugging task, as we have observed the positive correlation between *convergent thinking* and performance when they were reading and editing/checking the code in the code and console AOIs. *Cognitive load* was negatively correlated to the performance, and observed when students were reading and editing/checking the code and the questions, both in the initial and the later R-W episodes. In the solving and responding phases, *confusion* and *frustration* were again positively correlated to the performance, and only observed when students were reading the questions. On the other hand, *boredom* which was negatively correlated to the performance, was observed in the later R-W episodes in the code and the question AOIs, similar as in the initial R-W episodes. At last, contrary to our expectations, *delight* was negatively correlated to the performance in both, initial and later R-W episodes, when students were reading and making edits in the code.

In Table 7, we present how cognition and emotion co-exist among the students, which we have classified as low and high performing participants. The main difference in the initial R-W episodes is in the frequently displayed high levels of *attention*, *convergent thinking* and *frustration* by the high performing participants, compared to the *cognitive load*, *memory load*, and *boredom* displayed by the low performing participants. Once the students entered in the hypothesis verification loop, that is, the changing-testing phase, and the final stages of fixing the bugs, we have observed that they did not deviate from the behaviour they demonstrated in the encoding phase. The high performing

TABLE 6 Results from the linear regression based on later R-W episodes

Initial R-W episodes	AOI	Learning-related construct	Estimate	SE	t-Value	p-Value
		(Intercept)	0.57	0.5173	0.13	0.84
Later reading episode	Code	Cognitive load	−0.88	0.0076	−2.01	0.05
		Convergent thinking	1.35	0.0003	2.34	0.05
		Attention	1.57	0.0016	1.98	0.05
		Boredom	−0.97	0.0161	−2.06	0.05
		Delight	−0.72	0.0021	−2.02	0.05
	Question	Memory load	−1.68	0.0007	2.522	0.05
		Boredom	−1.46	0.0693	−2.54	0.05
		Confusion	1.98	0.0988	1.98	0.05
		Frustration	2.26	0.0103	−2.15	0.05
	Console	Convergent thinking	1.51	0.0010	3.003	0.01
Later writing episode	Code	Cognitive load	−2.61	0.0061	−2.18	0.05
		Attention	2.13	0.0006	2.58	0.01
		Convergent thinking	0.56	0.0018	1.98	0.05
		Delight	−0.62	0.0054	−1.99	0.05
	Question	Cognitive load	−3.08	0.0056	−2.61	0.01
		Attention	2.32	0.0005	2.59	0.01
		Boredom	−0.81	0.0039	−2.44	0.05
	Console	Convergent thinking	0.65	0.0013	1.973	0.05

TABLE 7 Significant variables as per R-W episodes compared to performance

	Code		Question		Console	
R-W episodes	Positive	Negative	Positive	Negative	Positive	Negative
Initial reading	Attention	Boredom	Attention	Memory load	Convergent thinking	Delight
	Convergent thinking	Delight	Convergent thinking	Boredom	Frustration	
	Frustration	Cognitive load				
Initial writing	Attention	Cognitive load	Convergent thinking	Cognitive load	Convergent thinking	
	Convergent thinking	Memory load		Boredom	Confusion	
	Frustration	Boredom				
		Delight				
Later reading	Attention	Cognitive load	Frustration	Memory load	Convergent thinking	
	Convergent thinking	Boredom	Confusion	Boredom		
		Delight				
Later writing	Attention	Cognitive load	Attention	Cognitive load	Convergent thinking	
	Convergent thinking	Delight		Boredom		

participants continued to display *convergent thinking* while reading and editing the code in the code and console AOIs, and *attention* while reading and editing the code or the questions. *Convergent thinking* was not significant when the students were reading and editing the questions during the later R-W episodes. The low performing participants continued to demonstrate *cognitive load* and *memory load* when reading or editing the code or the print statements in the questions, during the solving and the responding phases. These students also did not deviate from the *boredom state* that they have originally exhibited in the initial R-W episodes, while reading and editing the code or the questions.

5 | DISCUSSION AND CONCLUSION

5.1 | Interpretation of the results with respect to RQ1

The findings from our experiment demonstrate that particular cognitive-affective states are relevant and influential to both, the cognitive process of problem solving and the performance of students. Starting with the cognitive dimension, our findings demonstrate that *convergent thinking* and *attention* were positively

correlated to the performance, while *memory load* and *cognitive load* were negatively correlated to the performance for the entire code-debugging activity. Although these findings support the previous fundamental conclusions from instructional research regarding cognition and performance (Clark & Mayer, 2016; Mayer, 2002, 2003), our results provide new insights in the context of learning design for problem-solving activities in digital settings (Mayer, 1987, 1998).

Convergent thinking was an expected cognitive state in the encoding phase of problem solving, because during this phase, students usually recall specific problem-solving examples and check to see if they have a stored answer for the task at hand (Anderson, 1993; Tenison et al., 2016). This cognitive state is accompanied by emotions caused by pleasure or displeasure of performing a task (Shemyakina & Dan'ko, 2007). In our case, the low performing participants exhibited boredom, which suggests that this negative emotional induction caused decrease in the beta band (i.e., convergent thinking), which also led to decrease in the performance. Such situations can be downscaled if the learning activities and content are personalized to students' interests (Cordova & Lepper, 1996) and adapted to their proficiency (Mangaroska et al., 2019). *Memory load*, a task-centred dimension, was negatively correlated with the performance, and was observed in the initial reading of the questions and the small edits in the code, and later when students were reading the questions. One explanation might be that the design of the questions could have imposed load on students' working memory that caused adverse effects on the problem-solving performance for low performing participants. However, this might also be an indicator that the low performing participants had gaps in their domain-specific knowledge (i.e., knowledge in programming) or lacked knowledge of problem-solving methods. For example, there is a chance that the low performing participants had gaps in their syntax-based knowledge, because syntactic knowledge (i.e., the knowledge of how words can be combined in meaningful sentences, phrases, or utterances) is the only memory-related aspect in the code-debugging activity. In addition, syntactic knowledge is necessary for fixing bugs and for efficient coding in general (McCauley et al., 2008). Therefore, low-performing participants displaying high memory load is an opportunity for designing personalized content and scaffolds that would not cause adverse effects on the to-be-learned skills.

The implications from our method articulate the benefits of using MMLA as sufficiently sensitive technique to capture the complexities of cognitive engagement (compared to the long standing self- report measures; Greene, 2015), utilizing process-related data generated with sensors from the moment to moment tracking of clicks, facial expressions, gaze, and EEG activity of the brain. As advocated by Sinatra et al. (2015, p. 2), such grain-sized continuum at which engagement can be conceptualized, observed, and measured, ranges from the micro-level (i.e., individual in the moment, task, and learning activity) to the macro-level (e.g., group of learners in a class or a course), or from person-centred to context-centred, supporting the clarification for some of the various measurement and definition issues with the concept of engagement (Azevedo, 2015).

Considering the affective dimension, our findings demonstrate that *confusion* and *frustration* were positively correlated to the performance, while *delight* and *boredom* were negatively correlated to the performance for the entire code-debugging activity. *Frustration* was observed in the initial reading and editing of the code, while *confusion* was observed when students were checking the output in the console AOI based on the small edits in the code, and when reading the questions in the later R-W episodes. Frustration and confusion are natural and unavoidable states that learners demonstrate when engaged in deep learning (Meyer & Turner, 2006; Baker et al., 2010). In our case, we argue that frustration and confusion resulted from the 'checking' students did on the recalled problem-solving examples, when they became aware of the discrepancies between their knowledge and the problem in the task at hand, causing cognitive imbalance. Thus, as long as these states result from the cognitive processing and are not caused by an external stimulus, an intervention is not need it (Baker et al., 2010).

Boredom was present in the initial and later reading and editing of the code, but not when students were reading the output in the console AOI. We assume that the observed state of *boredom*, as a state of low arousal, was not cause because students were not challenged, but because these students (who also demonstrated low performance) might have felt being 'stuck' early on, which might have been caused by the gaps in their domain-specific knowledge (i.e., knowledge of programming) or due to the lack of problem-solving skills. At last, although *delight* has a positive valence and a high level of arousal, in our experiment we observed a negative correlation with the performance. This relation seems as counter-intuitive; however, we argue that *delight* might occur at the beginning of the problem solving (i.e., encoding stage) as a successful outcome from the initial small fixes in the code, which in turn might induce overconfidence in learners. This overconfidence can result in an overall slow progress, creating a negative effect on the performance.

Considering the effects from the affective, cognitive, and behavioural dimension of learning on the performance of students, our method has implications for educators and learning designers in the struggle to overcome the one-size-fits-all approach when it comes to developing and disseminating learning and assessment content (Gašević et al., 2015). Our method provides researchers and educators a set of multimodal data measures which have been extracted from related works (see Table 2) and allows them to account for students' cognitive, affective and behavioural processes. We agree that this list is not exhaustive, but provides a set of measures that are widely accepted in the learning technology literature, and offers certain implications. For example, it allows us to identify when students demonstrate behavioural engagement without strong cognitive engagement required for particular tasks. This is particularly important since behavioural engagement is often associated with assignments based on simple recall of lecture attendance, and is not a good indicator for achievement if higher order processing skills are expected to be developed. This can help educators to understand when and how students engage cognitively, and what affective and behavioural states complement particular cognitive states.

5.2 | Interpretation of the results with respect to RQ2

Expertise is one of the factors that affect student performance. In fact, modern educational psychology research suggests that learning outcomes result from the dynamic interaction of intra-individual factors, such as prior knowledge, motivation, cognition, emotions, and the contexts surrounding learners (Bronfenbrenner et al., 1998; National Academies of Sciences, 2018). Therefore, we have included expertise as one of the variables in our models to explain the debugging performance. Looking at the results shown in Table 4, one can notice that adding expertise (i.e., prior-knowledge) did not add much extra information to explain the variance in the performance. This result supports the decades of research work in educational psychology, which presents performance as a complex emulsion of potential (i.e., intra-individual factors) and opportunities (i.e., context). However, the models that are built using measures from multimodal data explain more than 70% of the variance in student performance, and thus, deserve more attention in future re- search. Such models can have profound implications for perspective inferences how students learn and what obstacles they face in computer-mediated learning environments, that instructors can utilize it to optimize instruction, content, and resources in relation to student potentials. Moreover, models build utilizing multimodal data can also aid to conceptualize, reveal, and measure constructs important for learning (e.g., mind wandering, Mills et al., 2020; convergent thinking, Razoumnikova, 2000), that could otherwise remain latent, but are important for students (not solely instructors) to become aware of their own capacities and behaviours. Yet, one key implication of these models is the support in the development of dynamic assessment models that can foster educational equality among students that have experienced different learning opportunities early in life (Alexander et al., 2009; Dumas et al., 2020), by not treating expertise as the most important factor in someone's performance, as well as include learning phenomena (e.g., intelligence (Thorndike, 1924), nonlinear improvement in student performance (Dumas & McNeish, 2017)) that educational psychologists have long struggled to study.

5.3 | Interpretation of the results with respect to RQ3

In the context of learning design, we posit that *cognitive load* and *memory load*, possibly caused by the discrepancy between the task processing demands and the processing capacities of the participants' who performed poorly (Paas & Van Merriënboer, 1994), should be *supervised and managed through tailored interventions early on in the problem-solving activity*. Managing *memory load* and *cognitive load* via actionable interventions (e.g., solutions with explanation, corrections of detected misconceptions), and encouraging active thinking by prompting learners to reason and to reflect (i.e., metacognitive awareness), can be seen as *promising approaches towards design of actionable feedback mechanisms* (e.g., *cognitive feedback*) that can prevent low performance (Van Merriënboer & Kirschner, 2017). Instead of focusing solely on corrective feedback (detection and correction of errors), instructors should be encouraged to consider *designing cognitive feedback* (i.e., feedback on the problem-solving process) *to stimulate learners to critically reflect and improve their metacognitive awareness* (Van Merriënboer & Kirschner, 2017; Hartman, 2001). Failure to teach metacognitive skills leads to 'vicious cycles' which occur when students are 'stuck' in certain states (e.g., boredom) and cannot move to more positive states (e.g., flow) during the learning activity. Our findings support the previous research (Hartman, 2001) that educators often focus on modelling cognition (i.e., how to perform a task) without modelling metacognition (i.e., how learners should think about and monitor their performance), demonstrating the many limits in education, in teaching students to be creative problem solvers (Mayer, 1998, Mayer, 1987).

Boredom and *delight* are states that also require to be managed early on in the learning activity and thereby, planned for in the learning design. Although humans are able to manage their emotions, the degree of success frequently depends on many factors, some of which are the situation, humans' management skills, and humans' temperament (Kagan, 1984). Hence, our recommendation to focus on *boredom* and *delight* early on is particularly important for learners who have lower domain-specific knowledge (in the particular topic) and skills to self-regulate their behaviour and emotions. Moreover, as previous research has shown *boredom* to be the most persistent and difficult to deal with emotional state (Baker et al., 2010), which also has profound effects on learning-related constructs such as productivity, engagement, performance, and stress (Gross & Muñoz, 1995), it should receive a greater research attention in the field of educational technology than any other affective state salient to learning.

Although *frustration* and *confusion* were observed to be positively correlated with the performance, these states need to be managed in the learning design to avoid negative cognitive loops and annoyance from being 'stuck' in learning situations for too long. In general, *frustration* and *confusion* are considered to be states that accompany deep learning if managed productively (Meyer & Turner, 2006; Dweck, 2002). For example, in our study we posit that *confusion* was demonstrated during the hypothesis verification stage when the students became metacognitively aware of discrepant events, and a sign of progress for the high performing participants that managed the state of confusion productively. Therefore, being aware that some level of *frustration* and *confusion* are critical for optimal learning (Craig et al., 2004; D'Mello et al., 2014; Lodge et al., 2018) and cannot be avoided in complex learning activities (D'Mello et al., 2014), suggests that these concepts salient to learning need to be considered in the learning design. Moreover, these insights also welcome design of interventions (with sufficient scaffolds to support learners to resolve the confusion) for particular groups of students, that can induce confusion to promote metacognitive awareness about the state of learners' knowledge, as suggested by Mandler (1990) interruption (discrepancy) theory.

5.4 | Limitations

Although our methods present the potential of multimodal data to capture constructs connected with learners' cognition and affective states during problem-solving activity, it is also subjected to certain limitations. First, we like to acknowledge the uniformity in student majors, suggesting that these findings do not apply to students from other majors. In other words, the problem solving was explored through a programming task, which requires domain-specific knowledge in computer science. The constructs connected with learners' cognition and affect might differ in other domains, such as the social sciences. Second, our students were performing the activity individually, which reflects a certain learning context. This means that the measurement of these constructs might vary in collaborative learning contexts. Third, our study privileged computer-mediated learning, where our students were engaged with a computing device for the entire activity. In future, we plan to explore the same constructs in a problem-solving activity that will require less focus on computing devices. Fourth, the study was performed in a controlled environment that might affect the ecology of the study, because the participants were aware of the physiological data collection, which may cause increase in desire 'to perform' and generate good biomarker metrics. Finally, to measure complex internal conditions within individuals emphasizing the natural sequence of events, we utilized objective multimodal data collected with sensors. Therefore, some might argue that other subjective measures generated from think aloud protocols or self- report measures such as invested mental effort, should be used in parallel, as they might shed light to outcomes different than the ones generated from the multimodal data. This limitation can be addressed in future studies when we can allow for the necessary disruptions (e.g., filling a questionnaire every 15 min) during a learning activity.

5.5 | Implications for the learning design and the development of learning systems

It has been widely acknowledged that learning in digital settings without an instructor or automated support that can distinguish between cognitive and affective difficulties, might cause an experience where students can easily give in to confusion, get frustrated or bored, and completely disengage from a learning activity (D'Mello & Graesser, 2012; Clark & Mayer, 2016; Van Merriënboer & Kirschner, 2017). In terms of potential implications for future research in the learning design and the development of learning systems, we posit that insights such as the ones we presented in this paper promise to support identification and modelling of variability in humans' natural behaviours. Some insights might be practical for designing educational interventions that intentionally puzzle learners to facilitate metacognitive awareness and teach the importance of reasoning and reflection during problem solving, compared to learning environments where students comfortably accumulate declarative knowledge without challenges (Hartman, 2001). Moreover, empirically verified multimodal proxies of cognitive and affective dimensions of problem solving can aid instructors to apply personalization of content and instruction enough to induce a learner to engage but not detract from the lesson. Other insights may advance automation in learning technologies to sense behaviours, cognitive and affective states, so that the system can adapt assessment (Mangaroska et al., 2019), interaction (Yoon & Narayanan, 2004), or regulate behaviour and emotions in humans through an interactive affect-support agents (Klein et al., 2002; Hone, 2006). In other words, the future research in design of learning technologies that are intuitive, easy to learn, and adaptable can leverage from users' knowledge, experience, and their engrained behavioural and language patterns (Oviatt, 2006). Furthermore, specific proxies from multimodal data promise to support the development of flexible and adaptive multimodal interfaces that can aid users in 'self-managing their cognitive load and minimizing related performance errors while solving complex real-world tasks' (Oviatt, 2006, p. 872). Such empirically verified proxies can shed light how, when, and why humans shift to more multimodal communication as their cognitive load increases during the different phases of information processing (Camp et al., 2001; Oviatt et al., 2004). This is especially important as future interfaces in education are expected to be designed to minimize cognitive load so that users can focus on the intrinsic difficulty of the tasks (e.g., in mathematics or programming), accommodating users' existing practices, and minimizing interruptions and distractions.

At last, going back to the contributions we provisionally established in the introduction, we like to posit that teaching declarative (i.e., knowing 'about') and procedural (i.e., knowing 'how-to') knowledge is not enough when teaching problem solving (Schraw, 1998), because this approach ignores the problem solver's affective states and interest in the problem (i.e., motivation). Hence, the instructional application is that the improvements in the learning design need to account not just for the physiological bases of cognition, but also the affective bases, as well as the social and cultural context of cognition (Mayer, 2003). This is what Van Merriënboer & Kirschner (2017) call holistic learning design that integrates declarative, procedural, metacognitive, and affective learning, and facilitates transfer of learning across contexts (i.e., application of previous knowledge and skills in new situations; Alexander & Winne, 2006).

ACKNOLEWDGEMENTS

This work was supported by the Research Council of Norway under the project FUTURE LEARNING (255129/H20). In addition, the authors are extremely grateful to the associate editor and the reviewers for their constructive comments and useful insights, which significantly improved the paper.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

ENDNOTES

[1] Generally, biomarker is anything that can be used as an indicator of some physiological state of an organism.

[2] Memory load is a task-centred dimension: the load imposed on working memory by the task (Sweller et al., 2019).

[3] Mental effort is a human-centred dimension: the amount of capacity or resources a learner allocates to accommodate the task demands (Sweller et al., 2019).

[4] Cognitive load is a multidimensional construct representing the level of perceived mental effort for thinking and reasoning while performing a particular task (Paas et al., 2003).

[5] https://www.neuroelectrics.com:3001/downloads/NEU$M_P.1_EN_2.7_EN$.pdf

[6] ICA is typically used for multiple source separation when there is a mixed signal (Xue et al., 2006). It is a powerful computational technique that divides the multisource signal into individual subcomponents on which further applications can be performed. ICA is also pertinent to blind source separation (BSS) or blind signal separation (Vorobyov & Cichocki, 2002), that is, when the source of a specific signal is not known (e.g., some noise in EEG data that is not from jaw movement).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12590.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Katerina Mangaroska https://orcid.org/0000-0002-7853-0429
Kshitij Sharma https://orcid.org/0000-0003-3364-637X

REFERENCES

Alexander, P. A., Schallert, D. L., & Reynolds, R. E. (2009). What is learning anyway? A topographical perspective considered. *Educational Psychologist*, *44*(3), 176–192. https://doi.org/10.1080/00461520903029006

Alexander, P. A., Winne, P. H., Corno, L., & Anderman, E. M. (2006). *Handbook of educational psychology*. 2nd. New York, NY: Routledge. https://doi.org/10.4324/9780203874790

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35–44. https://doi.org/10.1037/0003-066x.48.1.35

Anderson, J. R. (2013). *The architecture of cognition*. 1st. New York, NY: Psychology Press. https://doi.org/10.4324/9781315799438

Andreassi, J. L. (2010). *Psychophysiology: Human behavior and physiological response*. New York, NY: Psychology Press. https://doi.org/10.4324/9780203880340

Armato, A., Lanatà, A., & Scilingo, E. P. (2013). Comparitive study on photometric normalization algorithms for an innovative, robust and real-time eye gaze tracker. *Journal of Real-Time Image Processing*, *8*(1), 21–33. https://doi.org/10.1007/s11554-011-0217-6

Aula, A., & Surakka, V. (2002). Auditory Emotional Feedback Facilitates Human-Computer Interaction. In X. Faulkner, J. Finlay & F. Détienne (Eds.), *People and Computers XVI-Memorable Yet Invisible* (pp. 337–349). London, UK: Springer. https://doi.org/10.1007/978-1-4471-0105-5_20

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, *50*(1), 84–94. https://doi.org/10.1080/00461520.2015.1004069

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: issues and challenges. *Computers in Human Behavior*, *96*, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

Baker, R., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baltrušaitis, T., Robinson, P. & Morency, L. P. (2016). Openface: An open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1–10). https://doi.org/10.1109/WACV.2016.7477553.

Bednarik, R. (2012). Expertise-dependent visual attention strategies develop over time during debugging with multiple code representations. *International Journal of Human-Computer Studies*, *70*(2), 143–155. https://doi.org/10.1016/j.ijhcs.2011.09.003

Benedek, M., Schickel, R. J., Jauk, E., Fink, A., & Neubauer, A. C. (2014). Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia*, *56*, 393–400. https://doi.org/10.1016/j.neuropsychologia.2014.02.010

Bergdahl, N., Nouri, J., Fors, U., & Knutsson, O. (2020). Engagement, disengagement and performance when learning with technologies in upper secondary school. *Computers & Education*, *149*, 103783. https://doi.org/10.1016/j.compedu.2019.103783

Blikstein, P. (2013). Multimodal learning analytics. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (pp. 102–106). https://doi.org/10.1145/2460296.2460316

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Boekaerts, M. (2017). Cognitive load and self-regulation: Attempts to build a bridge. *Learning and Instruction*, *51*, 90–97. https://doi.org/10.1016/j.learninstruc.2017.07.001

Bronfenbrenner, U., et al. (1998). The ecology of developmental processes. In W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology*, (Vol. 1, pp. 993–1028). New York, USA: Wiley.

Camp, G., Paas, F., Rikers, R., & van Merrienboer, J. (2001). Dynamic problem selection in air traffic control training: a comparison between performance, mental effort and mental efficiency. *Computers in Human Behavior*, *17*(5-6), 575–595. https://doi.org/10.1016/s0747-5632(01)00028-0

Capdeferro, N., & Romero, M. (2012). Are online learners frustrated with collaborative learning experiences?. *The International Review of Research in Open and Distributed Learning*, *13*(2), 26. https://doi.org/10.19173/irrodl.v13i2.1127

Carp, J., & Compton, R. J. (2009). Alpha power is influenced by performance errors. *Psychophysiology*, *46*(2), 336–343. https://doi.org/10.1111/j.1469-8986.2008.00773.x

Chan, M. C. E., Ochoa, X., & Clarke, D. (2020). Multimodal learning analytics in a laboratory classroom. In M. Virvou, E. Alepis, G. Tsihrintzis & L. Jain (Eds.), *Machine learning paradigms* (pp. 131–156). Springer. https://doi.org/10.1007/978-3-030-13743-4_8

Chang, J. W., Wang, T., Lee, M. M., Su, C. & Chang, P. (2016). Impact of using creative thinking skills and open data on programming design in a computer-supported collaborative learning environment. In IEEE 16th international conference on advanced learning technologies (ICALT) (pp. 396–400). https://doi.org/10.1109/ICALT.2016.78.

Chen, C-M., & Wu, C-H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education*, *80*, 108–121. https://doi.org/10.1016/j.compedu.2014.08.015

Chen, S., & Epps, J. (2014). Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10.1080/07370024.2014.892428

Chen, S. J., & Caropreso, E. J. (2004). Influence of personality on online discussion. *Journal of Interactive Online Learning*, *3*(2), 1–17.

Chew, S. L., & Cerbin, W. J. (2021). The cognitive challenges of effective teaching. *The Journal of Economic Education*, *52*(1), 17–40. https://doi.org/10.1080/00220485.2020.1845266

Clarebout, G., & Elen, J. (2001). The ParlEuNet-project: problems with the validation of socio-constructivist design principles in ecological settings. *Computers in Human Behavior*, *17*(5-6), 453–464. https://doi.org/10.1016/s0747-5632(01)00019-x

Clark, R. C., Mayer, R. E., & Thalheimer, W. (2003). E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. *Performance Improvement*, *42*(5), 41–43. https://doi.org/10.1002/pfi.4930420510

Cooper, N. R., Burgess, A. P., Croft, R. J., & Gruzelier, J. H. (2006). Investigating evoked and induced electroencephalogram activity in task-related alpha power increases during an internally directed attention task. *NeuroReport*, *17*(2), 205–208. https://doi.org/10.1097/01.wnr.0000198433.29389.54

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730. https://doi.org/10.1037/0022-0663.88.4.715

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, *29*(3), 241–250. https://doi.org/10.1080/1358165042000283101

Csikszentmihalyi, M. (1996). *Creativity: Flow and the psychology of discovery and invention*. (p. 39). HarperCollins Publishers. https://digitalcommons.georgiasouthern.edu/ct2-library/35/

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, *38*(2), 105–134. https://doi.org/10.1007/s11251-009-9110-0

Department of Health. (2014). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* (Vol. 81). National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Di Mitri, D. (2019). Detecting medical simulation errors with machine learning and multimodal data. In 17th Conference on Artificial Intelligence in Medicine (pp. 1–6). Poznan, Poland.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099. https://doi.org/10.1037/a0032674

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016). Multi-sensor modeling of teacher instructional segments in live classrooms. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 177–184). https://doi.org/10.1145/2993148.2993158

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, *8*, https://doi.org/10.3389/fpsyg.2017.01153

Drachsler, H., & Schneider, J. (2018). JCAL Special Issue on Multimodal Learning Analytics. *Journal of Computer Assisted Learning*, *34*(4), 335–337. https://doi.org/10.1111/jcal.12291

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M. & Giannopoulos, I. (2018). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–13). https://doi.org/10.1145/3173574.3173856

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189x17725747

Dweck, C. S. (2002). Chapter 3 - Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. Aronson (Eds.), *Improving academic achievement* (pp. 37–60). Educational Psychology, Elsevier. https://doi.org/10.1016/B978-012064455-1/50006-3

Echeverria, V., Martinez-Maldonado, R. & Buckingham Shum, S. (2019). Towards collaboration translucence: giving meaning to multimodal group data. in proceedings of the 2019 chi conference on human factors in computing systems (p. 1-16). https://doi.org/10.1145/3290605.3300269

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion*, *19*(3), 313–332. https://doi.org/10.1080/02699930441000238

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gordon, E. M., Breeden, A. L., Bean, S. E., & Vaidya, C. J. (2014). Working memory-related changes in functional connectivity persist beyond task disengagement. *Human Brain Mapping*, *35*(3), 1004–1017. https://doi.org/10.1002/hbm.22230

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: reflections from over 20 years of research. *Educational Psychologist*, *50*(1), 14–30. https://doi.org/10.1080/00461520.2014.989230

Gross, J. J., & Muñoz, R. F. (1995). Emotion regulation and mental health. *Clinical Psychology: Science and Practice*, *2*(2), 151–164. https://doi.org/10.1111/j.1468-2850.1995.tb00036.x

Grunwald, M., Weiss, T., Krause, W., Beyer, L., Rost, R., Gutberlet, I., & Gertz, H-J. (1999). Power of theta waves in the EEG of human subjects increases during recall of haptic information. *Neuroscience Letters*, *260*(3), 189–192. https://doi.org/10.1016/s0304-3940(98)00990-2

Haapalainen, E., Kim, S., Forlizzi, J. F. & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In Proceedings of the 12th ACM International Conference on Ubiquitous Computing (pp. 301–310). https://doi.org/10.1145/1864349.1864395

Hartman, H. J. (2001). *Metacognition in learning and instruction: Theory, research and practice*, (Vol. 1), Neuropsychology and Cognition, Springer. https://doi.org/10.1007/978-94-017-2243-8

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, *26*(6), 1278–1288. https://doi.org/10.1016/j.chb.2010.05.031

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press. http://ukcatalogue.oup.com/product/9780199697083.do

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with Computers*, *18*(2), 227–245. https://doi.org/10.1016/j.intcom.2005.05.003

Jackson, D., Snow, E. R., & Corno, L. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of Educational Psychology*, (pp. 243–310). London, UK: Prentice Hall International.

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European Journal of Neuroscience*, *15*(8), 1395–1399. https://doi.org/10.1046/j.1460-9568.2002.01975

Johnson, S. D. (1997). Learning technological concepts and developing intellectual skills. In M. J. De Vries & A. Tamir (Eds.), *Shaping concepts of technology* (pp. 161–180). Springer. https://doi.org/10.1007/978-94-011-5598-4_13

Kagan, J. (1984). *The nature of the child*. New York, USA: Basic Books.

Kaller, C. P., Rahm, B., Bolkenius, K., & Unterrainer, J. M. (2009). Eye movements and visuospatial problem solving: Identifying separable phases of complex cognition. *Psychophysiology*, *46*(4), 818–830. https://doi.org/10.1111/j.1469-8986.2009.00821.x

Klein, J., Moon, Y., & Picard, R.W. (2002). This computer responds to user frustration. *Interacting with Computers*, *14*(2), 119–140. https://doi.org/10.1016/s0953-5438(01)00053-4

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., & Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neuroscience Letters*, *244*(2), 73–76. https://doi.org/10.1016/s0304-3940(98)00122-0

Kress, G., Jewitt, C., Ogborn, J., & Tsatsarelis, C. (2001). *Multimodal teaching and learning: The rhetorics of the science classroom*, (Vol. 72). London, UK: Bloomsbury.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

Li, Y., Chang, M., Kravcik, M., Popescu, E., Huang, R., Kinshuk, & Chen, N. S. (2016). *State-of-the-art and future directions of smart learning*. Lecture Notes in Educational Technology, (Vol. 1). Singapore: Springer. https://doi.org/10.1007/978-981-287-868-7

Lodge, J. M., Kennedy, G., Lockyer, L., Arguel, A., & Pachman, M. (2018). Understanding difficulties and resulting confusion in learning: an integrative review. *Frontiers in Education*, *3*. https://doi.org/10.3389/feduc.2018.00049

Malmberg, J., Järvelä, S., Holappa, J., Haataja, E., Huang, X., & Siipo, A. (2019). Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning?. *Computers in Human Behavior*, *96*, 235–245. https://doi.org/10.1016/j.chb.2018.06.030

Mandler, G. (1990). Interruption (discrepancy) theory: Review and extensions. In S. Fisher & C. L. Cooper (Eds.), *On the move: The psychology of change and transition* (Vol. 13, p. 32). Wiley & Sons Inc.

Mangaroska, K., Sharma, K., Giannakos, M., Trætteberg, H. & Dillenbourg, P. (2018). Gaze insights into debugging behavior using learner-centred analysis. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 350–359). https://doi.org/10.1145/3170358.3170386

Mangaroska, K., Vesin, B. & Giannakos, M. (2019). Elo-rating method: Towards adaptive assessment in e-learning. In IEEE 19th International Conference on Advanced Learning Technologies (ICALT) (Vol. 2161, pp. 380–382). https://doi.org/10.1109/ICALT.2019.00116

Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G. & Buckingham Shum, S. (2020a). From data to insights: A layered storytelling approach for multimodal learning analytics. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–15). https://doi.org/10.1145/3313831.3376148

Martinez-Maldonado, R., Echeverria, V., Schulte, J., Shibani, A., Mangaroska, K. & Shum, S. B. (2020). Moodoo: Indoor Positioning Analytics for Characterising Classroom Teaching. In Bittencourt I., Cukurova M., Muldner K., Luckin R., Millán E. (Eds.), *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science*, (Vol. *12163*). Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_29

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2019). Collocated collaboration analytics: principles and dilemmas for mining multimodal interaction data. *Human–Computer Interaction*, *34*(1), 1–50. https://doi.org/10.1080/07370024.2017.1338956

Martinez-Maldonado, R., Mangaroska, K., Schulte, J., Elliott, D., Axisa, C., & Shum, S. B. (2020c). Teacher tracking with integrity: What indoor positioning can reveal about instructional proxemics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), 1–27. https://doi.org/10.1145/3381017

Mayer, R. E. (1987). The elusive search for teachable aspects of problem solving. In J. A. Glover & R. R. Ronning (Eds.), *Historical foundations of educational psychology*. Perspectives on Individual Differences, (pp. 327–347). Boston, MA, USA: Springer. https://doi.org/10.1007/978-1-4899-3620-2_15

Mayer, R. E. (1998). *Instructional Science*, *26*(1/2), 49–63. https://doi.org/10.1023/a:1003088013286

Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation* (Vol. 41, pp. 85–139). Elsevier. https://doi.org/10.1016/S0079-7421(02)80005-6

Mayer, R. E. (2003). Chapter 3 - Memory and information processes. In I. Weiner, W. Reynolds & G. Miller (Eds.), *Handbook of Psychology*, Part Two. Cognitive Contributions to Learning, Development, and Instruction, (Vol. 7, pp. 47–57). Hoboken, NJ: John Wiley and Sons, Inc. https://doi.org/10.1002/0471264385.wei0703

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. Berliner & R. Calfee, *Handbook of Educational Psychology*, (pp. 47–62). New York, USA: Routledge. Taylor & Francis Group.

McCauley, R., Fitzgerald, S., Lewandowski, G., Murphy, L., Simon, B., Thomas, L., & Zander, C. (2008). Debugging: A review of the literature from an educational perspective. *Computer Science Education*, *18*(2), 67–92. https://doi.org/10.1080/08993400802114581

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K. & Graesser, A. (2007). Facial features for affective state detection in learning environments. In McNamara D. S. & Trafton J. G. (Eds.), Proceedings of the 29th Annual Meeting of the Cognitive Science Society (Vol. 29).

Meyer, D. K., & Turner, J. C. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, *18*(4), 377–390. https://doi.org/10.1007/s10648-006-9032-1

Mills, C., Bosch, N., Krasich, K. & D'Mello, S. K. (2019) Reducing mind-wandering during vicarious learning from an intelligent tutoring system. In Isotani S., Millán E., Ogan A., Hastings P., McLaren B., Luckin R. (Eds.), Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, (Vol. 11625). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_25

Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (2021). Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, *36*(4), 306–332. https://doi.org/10.1080/07370024.2020.1716762

Mousavinasab, E., Zarifsanaiey, N. R., Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257

National Academies of Sciences & Medicine (2018). How people learn II: Learners, contexts, and cultures.

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G. & Castells, J. (2018). The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 360–364). https://doi.org/10.1145/3170358.3170406

Olsen, Anneli (2012) The Tobii I-VT Fixation Filter. Algorithm description. Tobii Technology.

Oviatt, S. (2006). Human-centered design meets cognitive load theory: Designing interfaces that help people think. In Proceedings of the 14th ACM International Conference on Multimedia (pp. 871–880). https://doi.org/10.1145/1180639.1180831

Oviatt, S., Coulston, R. & Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. In Proceedings of the 6th International Conference on Multimodal Interfaces (pp. 129–136). https://doi.org/10.1145/1027933.1027957

Oviatt, S., Grafsgaard, J., Chen, L., & Ochoa, X. (2018). Multimodal learning analytics: Assessing learners' mental state during the process of learning. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos & A. Krüger (Eds.), *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition* (pp. 331–374). Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/3107990.3108003

Oviatt, S., Schuller, B., Cohen, P., Sonntag, D., Potamianos, G., & Krüger, A. (2017). *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations*. 1, Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/3015783

Paas, F., Renkl, Al., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/s15326985ep3801_1

Paas, F. G. W. C., Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371. https://doi.org/10.1007/bf02213420

Panitz, T. (1999). *The case for student centered instruction via collaborative learning paradigms*. U. S. Department of Education.

Pijeira-Díaz, H. J., Drachsler, H., Kirschner, P. A., & Järvelä, S. (2018). Profiling sympathetic arousal in a physics course: How active are students?. *Journal of Computer Assisted Learning*, *34*(4), 397–408. https://doi.org/10.1111/jcal.12271

Prieto, L. P., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M. J., & Dillenbourg, P. (2018). Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning*, *34*(2), 193–203. https://doi.org/10.1111/jcal.12232

Razoumnikova, O. M. (2000). Functional organization of different brain areas during convergent and divergent thinking: an EEG investigation. *Cognitive Brain Research*, *10*(1-2), 11–18. https://doi.org/10.1016/s0926-6410(00)00017-3

Ritella, G., & Hakkarainen, K. (2012). Instrumental genesis in technology-mediated learning: From double stimulation to expansive knowledge practices. *International Journal of Computer-Supported Collaborative Learning*, *7*(2), 239–258. https://doi.org/10.1007/s11412-012-9144-1

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

Santhanam, R., Sasidharan, S., & Webster, J. (2008). Using self-regulatory learning to enhance e- learning-based information technology training. *Information Systems Research*, *19*(1), 26–47. https://doi.org/10.1287/isre.1070.0141

Schneider, B. (2020). A methodology for capturing joint visual attention using Mobile eye-trackers, a methodology for capturing joint visual attention using Mobile eye-trackers. *Journal of Visualized Experiments (Jove)*, (155), e60670. https://doi.org/10.3791/60670.

Schraw, G. (1998). *Instructional Science*, *26*(1/2), 113–125. https://doi.org/10.1023/a:1003044231033

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). "With-me-ness": A gaze-measure for students' attention in MOOCs. In Proceedings of International Conference of the Learning Sciences 2014 (pp. 1017–1022). https://doi.org/10.22318/icls2014.1017

Sharma, K., Mangaroska, K., Giannakos, M., & Dillenbourg, P. (2018). Interlacing Gaze and Actions to Explain the Debugging Process. In J. Kay & R. Luckin (Eds.), *Rethinking Learning in the Digital Age: Making the Learning Sciences Count*. In Proceedings of the 13th International Conference of the Learning Sciences (ICLS), (vol. 1). London, UK: International Society of the Learning Sciences. https://doi.org/10.22318/cscl2018.640

Shemyakina, N., & Dan'ko, S. (2007). Changes in the power and coherence of the β 2 EEG band in subjects performing creative tasks using emotionally significant and emotionally neutral words. *Human Physiology*, *33*(1), 20–26. https://doi.org/10.1134/S0362119707010033

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*, *12*(2), 176–189. https://www.jstor.org/stable/jeductechsoci.12.2.176

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Spering, M., Wagener, D., & Funke, J. (2005). BRIEF REPORT. *Cognition & Emotion*, *19*(8), 1252–1261. Psychology Press. Taylor & Francis Group. https://doi.org/10.1080/02699930500304886

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366–377. https://doi.org/10.1111/jcal.12263

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, *24*(18), 2174–2180. https://doi.org/10.1016/j.cub.2014.07.066

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Taub, M., & Azevedo, R. (2019). How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive srl processes during learning with an intelligent tutoring system? *International Journal of Artificial Intelligence in Education*, *29*(1), 1–28. https://doi.org/10.1007/s40593-018-0165-4

Tenison, C., Fincham, J. M., & Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive Psychology*, *87*, 1–28. https://doi.org/10.1016/j.cogpsych.2016.03.001

Thorndike, E. L. (1924). Measurement of Intelligence. *Psychological Review*, *31*(3), 219–252. https://doi.org/10.1037/h0073975

Trætteberg, H., Mavroudi, A., Giannakos, M., & Krogstie, J. (2016). Adaptable learning and learning analytics: A case study in a programming course. In *European conference on technology enhanced learning* (pp. 665–668). Springer, Cham. https://doi.org/10.1007/978-3-319-45153-4_87

Van Merriënboer, J. J., & Kirschner, P. A. (2017). *Ten steps to complex learning: A systematic approach to four-component instructional design*. New York, USA: Routledge. https://doi.org/10.4324/9781315113210

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*(3), 209–249. https://doi.org/10.1207/S1532690XCI2103_01

Vorobyov, S., & Cichocki, A. (2002). Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis. *Biological Cybernetics*, *86*(4), 293–303. https://doi.org/10.1007/s00422-001-0298-6

Wang, C., Fang, T., & Miao, R. (2018). Learning performance and cognitive load in mobile learning: Impact of interaction complexity. *Journal of Computer Assisted Learning*, *34*(6), 917–927. https://doi.org/10.1111/jcal.12300

Wang, M., Wu, B., Chen, N.-S., Spector, J. M., et al. (2013). Connecting problem-solving and knowledge- construction processes in a visualization-based learning environment. *Computers & Education*, *68*, 293–306. https://doi.org/10.1016/j.compedu.2013.05.004

Xue, Z., Li, J., Li, S. & Wan, B. (2006). Using ICA to remove eye blink and power line artifacts in EEG. In First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06) (Vol. 3, pp. 107–110). https://doi.org/10.1109/ICICIC.2006.543.

Yoon, D. & Narayanan, N. H. (2004). Mental imagery in problem solving: An eye tracking study. In Proceedings of the 2004 Symposium on Eye

Tracking Research & Applications (ETRA '04) (pp. 77–84). https://doi.org/10.1145/968363.968382

Zhang, W., Sjoerds, Z., & Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage*, *210*, 116572. https://doi.org/10.1016/j.neuroimage.2020.116572

Zhou, Z., Hu, L., Sun, C., Li, M., Guo, F., & Zhao, Q. (2019). The effect of zhongyong thinking on remote association thinking: An EEG study. *Frontiers in Psychology*, *10*. 1–9. https://doi.org/10.3389/fpsyg.2019.00207

How to cite this article: Mangaroska, K., Sharma, K., Gašević, D., & Giannakos, M. (2022). Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity. *Journal of Computer Assisted Learning*, *38*(1), 40–59. https://doi.org/10.1111/jcal.12590