Multimodal Learning Analytics to Inform Learning Design: Lessons Learned from Computing Education

Katerina Mangaroska[1], Kshitij Sharma[2], Dragan Gašević[3], Michalis Giannakos[4]

Abstract

Programming is a complex learning activity that involves coordination of cognitive processes and affective states. These aspects are often considered individually in computing education research, demonstrating limited understanding of how and when students learn best. This issue confines researchers to contextualize evidence-driven outcomes when learning behaviour deviates from pedagogical intentions. Multimodal learning analytics (MMLA) captures data essential for measuring constructs (e.g., cognitive load, confusion) that are posited in the learning sciences as important for learning, and cannot effectively be measured solely with the use of programming process data (IDE-log data). Thus, we augmented IDE-log data with physiological data (e.g., gaze data) and participants' facial expressions, collected during a debugging learning activity. The findings emphasize the need for learning analytics that are consequential for learning, rather than easy and convenient to collect. In that regard, our paper aims to provoke productive reflections and conversations about the potential of MMLA to expand and advance the synergy of learning analytics and learning design among the community of educators from a post-evaluation design-aware process to a permanent monitoring process of adaptation.

Notes for Practice

- Multimodal learning analytics gathers insights into learners' actual behaviour and their cognitive-affective states, expanding the context and impact of learning analytics research by posing new methods and techniques to move forward the synergy between learning analytics and learning design.

- Measures from multimodal data can expand learning analytics that have the capacity to align with learning design and learning theories from the learning sciences, are consequential for learning, and can inform and influence the design (or its refinement) of learning activities and objectives in digital settings.

- Use of predictive modelling based on combinations of multimodal data streams affords generation of relevant learning analytics measures and showcases how the alignment between learning analytics and learning design can be achieved in computing education context.

Corresponding author [1]*Email:* mangaroska@ntnu.no *Address: Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway, ORCID ID:* https://orcid.org/0000-0002-7853-0429
[2]*Email:* kshitij.sharma@ntnu.no *Address: Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway, ORCID ID:* https://orcid.org/0000-0003-3364-637X
[3]*Email:* dragan.gasevic@monash.edu *Address: Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia, ORCID ID:* https://orcid.org/0000-0001-9265-1908
[4]*Email:* michailg@ntnu.no *Address: Department of Computer Science, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway, ORCID ID:* https://orcid.org/0000-0002-8016-6208

1. Introduction

Ten years of research in learning analytics (LA) has generated valuable data-driven approaches that assist educators to comprehend and evaluate their efforts and innovations in the design of learning activities (Lockyer & Dawson, 2012). However, data-driven approaches are not sufficiently informative for explaining inconsistencies, avoiding misconceptions, and clarifying contextual conditions that affect learning (Reimann, 2016). In fact, theoretical grounding is essential to

identifying relevant associations between LA and learning-oriented phenomena (Gašević, Kovanović, & Joksimović, 2017; Neelen & Kirschner, 2017). Although attempts to connect LA with the learning sciences (LS) is emphasized by various researchers (Neelen & Kirschner, 2017; Reimann, 2016), the role of theory in this connection is underexplored, both in terms of creating effective learning designs (LDs), as well as informing measures and models used in LA. To that end, we aim to emphasize the **need for LA that are consequential for learning, rather than easy and convenient to collect, as well as aligned with LD and learning theories from the LS** for which there is adequate empirical evidence (Neelen & Kirschner, 2017; Reimann, 2016).

Computing education (CE) is an area in which the use of LA to inform LD offers much promise (Blikstein et al., 2014; Hundhausen, Olivares, & Carter, 2017), as programming allows us to capture fine-grained digital traces of learners (Blikstein et al., 2014) and is a skill that requires mastering proper teaching strategies and support (Kurland, Pea, Clement, & Mawby, 1986). Programming is more than just the capability to generate code; it requires application of problem-solving in open-ended, unbounded solution spaces (Marion, Impagliazzo, St. Clair, Soroka, & Whitfield, 2007). Educators teach programming through authentic learning tasks based on practical educational approaches (e.g., problem-solving), which are considered essential for learning (Van Merriënboer, Kirschner, & Kester, 2003). Thus, as a problem-solving activity, programming entails many behavioural and cognitive steps, governed by metacognitive awareness and emotions, to accomplish the goal of the task at hand (Beattie, Collins, & McInnes, 1997; Dörner & Funke, 2017). For example, when solving complex problems, learners need to alter many interrelated variables, some of which cause cognitive disequilibrium (D'Mello & Graesser, 2012) that intertwine various behavioural (e.g., attention), cognitive (e.g., memory load), and affective (e.g., frustration) learning dimensions. This is also shown in the work of Snow, Corno, and Jackson III (1996) who hypothesize that cognition, motivation, and emotions are the three fundamental components of learning, and as such, all should be considered when designing learning activities that facilitate deep learning (Beattie et al., 1997).

Whilst important achievements have been obtained in the last decade in CE using LA (Blikstein et al., 2014; Hundhausen et al., 2017), most of the work focuses on modelling and predicting behaviour, and strategies from *programming process data (i.e., IDE-log data)*. This approach helped researchers to understand students' compilation behaviour and design accurate scaffolds (Jadud, 2006), explore differences between novices and experts (Blikstein et al., 2014) to improve the pedagogical content knowledge (Shulman, 1986), and distinguish various programming strategies students employ to arrive at their solutions (Bruce et al., 2004; Perkins, Hancock, Hobbs, Martin, & Simmons, 1986; Turkle & Papert, 1992). As a result, the connection between CE and LD has not been sufficiently established (O'Grady, 2012), first, because most of the research in LA and CE has been focused on ways that LA can be used to support computer programming (Hundhausen et al., 2017; Ihantola et al., 2015), and second, because CE is a relatively young discipline, very dynamic, and highly interdisciplinary.

To emphasize the importance of LD in CE (Schulte, Magenheim, Müller, & Budde, 2017), we elaborated on the potential of multimodal learning analytics (MMLA) to capture data essential for measuring constructs (e.g., cognitive load, confusion, frustration, selective attention) that are posited in the LS as important for learning, and cannot effectively be measured solely with the use of programming process data (i.e., IDE-log). We also focused on how MMLA can be used to understand how effectively students use the opportunities for learning as given in the LD. This exploration assumes the existence of comprehensive data generated at various levels (Hundhausen et al., 2017), which can inform and influence the design (or its refinement) of learning activities and objectives, particularly in digital settings. For example, if a learner is disengaging because there is an increase in cognitive demand, compared to a disengagement situation when a learner is frustrated or bored, a proactive intervention should consider different approaches (e.g., explaining misconceptions versus challenging the user). The MMLA approach gathers insights into learners' actual behaviour (not recalled behaviour or subjective impressions as self-reported data does) and their cognitive-affective states, *expanding the context and impact of LA research by posing new methods and techniques to move forward the synergy between LA and* **LD** (Blikstein & Worsley, 2016; Ochoa & Worsley, 2016; Oviatt, Grafsgaard, Chen, & Ochoa, 2018). To that end, MMLA provides significant insights for LDs based on learner behaviours that consider the temporal dimension (associated with natural frequencies of behaviour and emotions), as well as strategies employed from the LS (e.g., affective learning). Hence, we aim to extend the discussion and provide practical guidelines by addressing the following research questions:

RQ1: What insights can we derive from MMLA that can augment traditional LA with respect to LD in the context of CE?

RQ2: What implications can MMLA provides to advance the alignment between LA and LD in CE?

Consequently, we designed an open-ended learning activity (i.e., debugging task) that requires the application of problem-solving via code generation and adjustment. The IDE-log data was augmented with physiological data (i.e., gaze, electrodermal activity [EDA], heart rate, BVP, and skin temperature) and participants' facial expressions. The findings were used to interpret learners' actual behaviours with respect to potential LD guidelines based on learning theories from the LS for which

there is adequate empirical evidence, and selection of LA (informed by theories and grounded in evidence) that are relevant and consequential for learning.

2. Background

2.1. Learning Analytics for Learning Design

Learning analytics (LA) is an interdisciplinary field that embraces a holistic approach to study learning and learning contexts, and to address questions in educational research (Gašević, Dawson, & Siemens, 2015). However, understanding and optimizing learning involves more than just collecting and processing data (Gašević et al., 2017); it also tackles theory and learning design activities. For example, educators need to decide *how to (re)design the teaching–learning process* when *learning behaviour deviates from the pedagogical intention*, and *frame aspects of evidence-driven recommendations within a socio-technical system* (Wise & Jung, 2019). Thus, aligning LA with LD emphasizes the idea that "human decision-making and consequent actions are as much a part of any successful analytics solution as its technical components" (Van Harmelen & Workman, 2012, p. 4).

Lockyer and Dawson (2012) were among the first researchers to highlight the potential of integrating LA and LD, to increase the understanding of student behaviour and their engagement with different LDs. The ideas of Lockyer and Dawson (2012) have generated research at the intersection of LA and LD that produced several conceptual frameworks (Bakharia et al., 2016; Hernández-Leo, Martinez-Maldonado, Pardo, Muñoz-Cristóbal, & Rodríguez-Triana, 2019; Persico & Pozzi, 2015; Rienties, Nguyen, Holmes, & Reedy, 2017). Bakharia et al. (2016) proposed a framework with five LA dimensions that can support educators with evaluations of their LDs, and highlighted their role in bringing context to data interpretation. The main benefit of this approach where the focus is in the pedagogical intent and centred around the instructor, is that the selected LA will be useful and meaningful to the instructor. However, this approach overlooks other relevant LA for which instructors are ignorant. Similarly, the work of Hernández-Leo et al. (2019) extends the role of educators as designers of learning activities, and presents a framework of layers (i.e., LA, design analytics, and community analytics) as guidelines for how to leverage the benefits of LA into the LD space. Others, such as Persico and Pozzi (2015) focused on representations and tools to connect LA with LD, while Rienties et al. (2017) investigated patterns in a ten-year data set to understand what and how LD elements affect the engagement, behaviour, and performance of students. The main benefit of Rienties et al.'s (2017) approach is the focus on the overall patterns in the data, rather than on the needs instructors have from the generated data. This way, researchers might discover relevant associations that are latent and difficult to notice because they are not covered by pedagogical objectives. Nonetheless, all the proposed frameworks emphasize the potentially valuable integration of LA with LD; however, the content of these frameworks is presented at highly descriptive levels, making any meaningful operationalization difficult (Corrin et al., 2016).

2.2. Multimodal Learning Analytics Research

Most of the research in LA thus far has focused on computer-mediated learning (Mangaroska & Giannakos, 2018) and highly structured learning activities (Blikstein & Worsley, 2016). However, learning is multimodal in nature and not restricted solely to interactions with technology or in front of a computer screen (Kress, 2001). Therefore, we need innovative approaches that education researchers can use to generate more nuanced understanding of the learning process and its complexities (Blikstein & Worsley, 2016; Ochoa & Worsley, 2016; Oviatt et al., 2018). One such example is MMLA, which extends the context and application of LA, by considering not only activity patterns (e.g., user behaviour from click-stream data), but also learners' physiological and neural patterns (e.g., gaze data, brain activity data; Oviatt et al., 2018).

As a term, MMLA does not have a generally accepted definition, but for the purpose of this paper, we will use the following: "a set of techniques employing multiple sources of data (video, logs, text, artifacts, audio, gestures, biosensors) to examine learning in realistic, ecologically valid, social, mixed-media learning environments" (Blikstein & Worsley, 2016, p. 233). So far, insights generated with MMLA have opened a wide array of research opportunities to challenge previously held assumptions about learning (e.g., better to be frustrated than bored; Baker, D'Mello, Rodrigo, & Graesser, 2010), which coupled with theory, can facilitate research towards creating novel assessment methods (Worsley & Blikstein, 2013), feedback (Echeverria, Martinez-Maldonado, & Buckingham Shum, 2019; Ochoa et al., 2018), and teaching–learning practices (Martinez-Maldonado et al., 2020; Rodríguez-Triana, Prieto, Martínez-Monés, Asensio-Pérez, & Dimitriadis, 2018), or learning environments (Melero, Hernández-Leo, Sun, Santos, & Blat, 2015; Ogan, 2019) that correspond to the current technological and sociocultural advancements.

At present, research in MMLA has focused on studying and modelling learning strategies (Mangaroska, Sharma, Giannakos, Træteberg, & Dillenbourg, 2018; Worsley & Blikstein, 2015), predicting high-level constructs such as learners' attention and engagement (Chan, Ochoa, & Clarke, 2020), building more accurate learner models (Giannakos, Sharma, Pappas,

Kostakos, & Velloso, 2019), designing multimodal learning interfaces (Echeverria et al., 2019), or generating new insights into teaching and learning at a more fine-grained level (Martinez-Maldonado et al., 2020; Prieto, Sharma, Kidzinski, Rodríguez-Triana, & Dillenbourg, 2018). In fact, MMLA promises to bridge theory and complex learning behaviour (Worsley, 2014), support educators' role in addressing students' individual needs, expectations, and skills in physical (Ogan, 2019) and digital (Ochoa et al., 2018) learning settings, as well as tackle the complexities of orchestrating (e.g., designing, managing, adapting) learning activities at multiple social levels beyond the challenges (e.g., time, mental effort) instructors face in their everyday practice (Prieto et al., 2018).

Although these previously mentioned studies demonstrated methods and techniques to deconstruct learning at fine levels of detail, and derived more comprehensive and nuanced insights into learner behaviours, actions, and states, only few at present (Echeverria et al., 2019; Rodríguez-Triana et al., 2018) touched upon the importance of aligning learning analytics with learning design, considering the applicability of these novel techniques in learning and teaching. Consequently, we add to the body of knowledge how that alignment can be further advanced by harnessing MMLA, and how the complexity of this approach is congruous with learning theories and their practical application in education, as well as beneficial to be studied at scale.

2.3. Multimodal Learning Analytics Research in Computing Education

In computing education, students spend most of their time working on programming assignments in integrated development environments (IDE). Thus, it is logically educators and researchers who use IDEs to collect continuous streams of data (i.e., IDE-log data) in order to make sense of learners' actions and behaviours, and improve the processes of teaching and learning programming (Altadmri & Brown, 2015; Hundhausen et al., 2017). Nowadays, the functionality of many IDEs can be extended with plugins that collect data with additional features, lifting the previous technological restrictions (Hundhausen et al., 2017). Examples of such research are asynchronous discussion posts (Carter & Hundhausen, 2015), automatic testing against test cases (Edwards & Perez-Quinones, 2008), survey/quiz data to get insights into learner attitudes and conceptual understanding of tasks (Ihantola, Sorva, & Vihavainen, 2014), studies collecting data from eye trackers (Busjahn et al., 2014; Kevic et al., 2015; Mangaroska et al., 2018), mouse and keyboard pressure (Arapakis, Lalmas, & Valkanas, 2014; Begel, 2016), heart rate (Ahonen et al., 2016), and electro-dermal activity (Müller, 2015). Although some of the studies tried to harness MMLA (usually utilizing one modality), none captured cognitive, affective, and behavioural states of learners simultaneously, nor explored programming as a problem-solving activity.

Multimodal learning analytics does not yet have wide application in computing education (Hundhausen et al., 2017). One early example is the study by Parnin (2011), who used electromyography (EMG) to measure programmers' sub-vocal utterances and found that "subvocalization was strongly associated with making edits to the code" (p. 200). Later, Fritz, Begel, Müller, Yigit-Elliott, and Züger (2014) used data from gaze, electroencephalographic (EEG) activity, and EDA, to classify the difficulty of code comprehension tasks for the purpose of developing programming support tools. Crk and Kluthe (2014) explored alpha and theta band frequencies and found that a decrease of alpha activity, and often an increase in theta activity, reflects an increase in cognitive load. Müller (2015) used eye tracking, EEG, and physiological data to assess the change in programmers' emotions while coding and to model their perceived difficulty. Lastly, Grover et al. (2016) used data from gaze, gesture, posture, and body movement to develop a computational model that can automatically detect collaboration in collaborative programming scenarios.

All these studies show that data from multiple modalities, captured when learners work within unbounded solution spaces, can generate rich information about behaviour and cognition, while accounting for the unique artifacts that learners develop. None have discussed how these findings can be translated in applicable design guidelines to improve the design of learning activities, instructional methods for teaching particular skills/concepts, and overall course redesign. In that respect, there is limited research discussing or exploring a co-occurrence of cognitive and affective states in learners in CE, and what significant features this co-occurrence brings that promise to support versatile improvements in the learning design (e.g., setting the right feedback loop, design of interventions, modelling learning strategies) in digital learning settings.

3. Methodology

3.1. Approach

The baseline of our analysis is *a model (i.e., M1) that solely relies on programming process data (i.e., IDE-log data)* and can easily be compared with MMLA-models (i.e., M2–M8) that include sensor-data coming from various devices (e.g., eye tracker, wristband, camera). Our baseline (i.e., base model) is inspired by the work of Turkle and Papert (1992) and Sharma, Mangaroska, Trætteberg, Lee-Cultura, and Giannakos (2018), who uncovered different programming strategies (i.e., methods of solving programming problems) among learners using the number of test runs and the improvements in unit test success.

Their reported strategies include "thinkers" (low number of main method runs, but high time difference between two runs) and "probers" (high number of main method runs, but low time difference between two runs). The base model in this study captures the activity of learners in terms of how much time they allocate to *understanding the code (i.e., reading)* and *fixing the bug (i.e., writing)*. Moreover, it also includes *prior knowledge (i.e., expertise)* and the *use of IDE support tools* for debugging (e.g., DebugView).

For the study, we designed and implemented a debugging activity[1] for computer science (CS) students (i.e., participants). The main task was debugging a Java class, named Person, that manages parent–child relationships (Mangaroska et al., 2018). The provided code tried, but failed, to ensure consistent object relationships, such as the mother of a child is female. The participants could check the correctness of the code by running the provided print statements in the main method of the code.

Next, we deconstructed the debugging process in three main stages: comprehension, changing-testing, and fixing. For each of the stages, we examined the IDE-log data, the fixation-duration on the defined Areas of Interest (AOIs; Figure 1), the two-way transitions between the various AOIs, and the rest of the features that could be automatically extracted from eye, skin, and facial data (Table 2).

3.2. Participants and Procedure

During Spring 2019, an experiment was performed at a contrived computer lab setting at the Norwegian University of Science and Technology with 46 CS majors (8 females and 38 males). The mean age of the participants was 21.87 years (Std. Dev. = 1.44 years). All participants had already taken a Java course where they were predominantly using Eclipse[2].
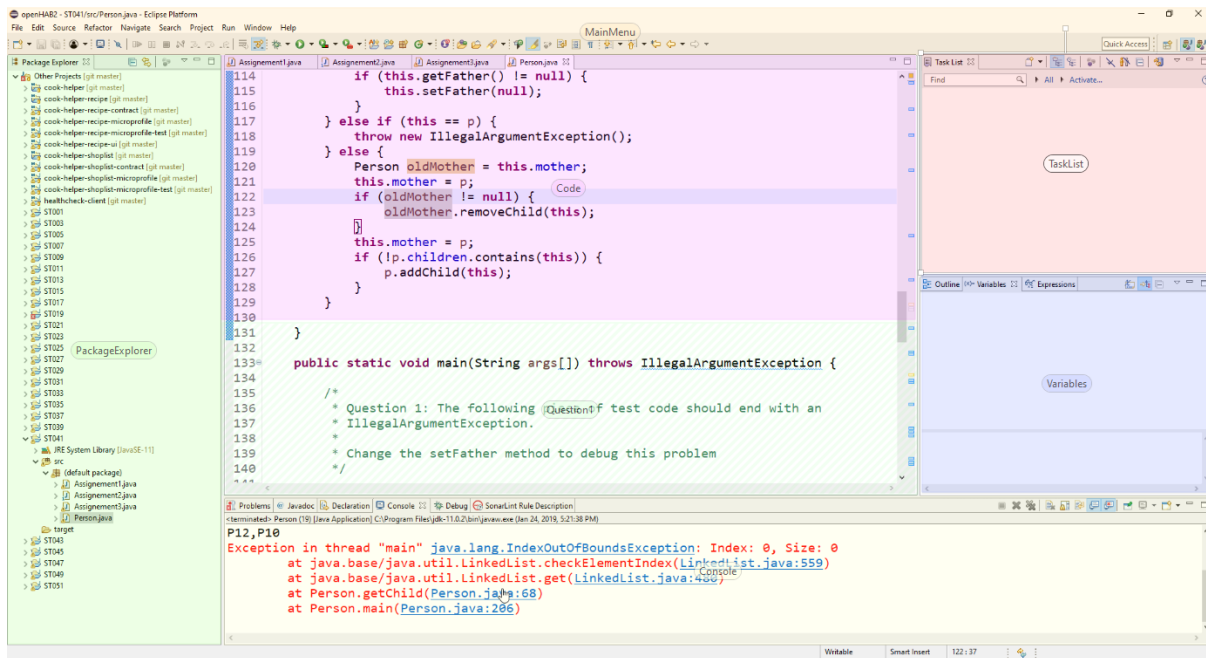

Figure 1. AOIs in Eclipse IDE

Upon arrival in the lab, the participants signed an informed consent form, after which the lead researcher placed a sensor device (i.e., Empatica E4) on the participant's wrist. Prior to the debugging tasks, each participant was asked to pass an automatic calibration routine. The wristband data streams were calibrated using the in-built calibration from the Empatica mobile app, while the eye tracker was calibrated using a 5-point calibration process.

Next, the participants were asked to finish three small debugging tasks (easy, medium, difficult) within 20 minutes. This pre-task test was used to decide the participants' expertise. After this, the participants were given 40 minutes to solve the main task, accompanied by five debugging sub-tasks (i.e., questions) presented as a part of the main method. The main task was cognitively challenging, requiring participants to maintain sustained attention and engagement. The steps from the study are shown in Figure 2.

---

[1] Debugging consists of four stages: understanding the code, testing the code, locating the bugs, and fixing the bugs.
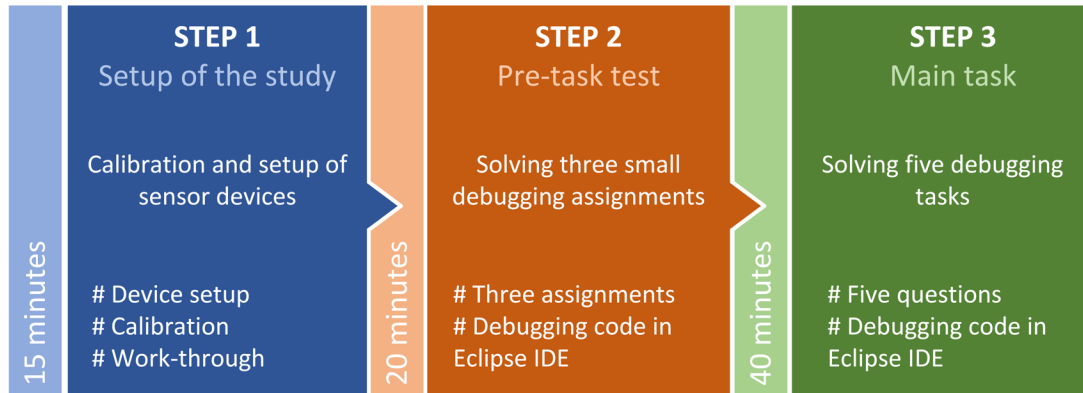[2] https://www.eclipse.org/ide/

Figure 2. The steps of the debugging study

3.3. Data Collection

Data was collected from four main sources: eye tracker (i.e., gaze data), wristband sensor (i.e., physiological data: HR, BVP, temperature, and EDA), camera (i.e., video data with participants' faces), and IDE-logs (i.e., click-streams).

Eye-tracking data. To record participants' gaze, we used a Tobii X3-120 eye-tracking device at 120 Hz sampling rate and using 5-point calibration. The device is non-invasive and mounted at the bottom of a computer screen. The resolution was 1920x1080 and participants' distance from the screen was 50–70 cm.

Physiological data from a wristband sensor. To record arousal data, we used the Empatica E4 wristband sensor. Four different measurements were recorded: 1) heart rate at 1 Hz, 2) EDA at 64 Hz, 3) body temperature at 4 Hz, and 4) blood volume pulse at 4 Hz.

Video data. To capture facial expressions, the authors used a LogiTech web camera, pointed straight at the participants' faces from the computer screen, capturing video at 30 frames-per-second (FPS). The web camera focus zoomed at 150% onto their faces. During the task, participants exhibited minimal body and gesture interaction; hence, the video recordings held high-quality data from participants' facial expressions. The video resolution was 640x480.

Log data. An Eclipse plugin (Mangaroska et al., 2018), was used to gather the reading and writing behaviour. This plugin captures the state of the program every time a student saves the program, either by clicking on the "save" button or by pressing "CTRL+S."

3.4. Data Pre-Processing

Biosensor recordings contain artifacts as a result of excessive hand movement, adjustments of sensors, variations in the pressure between the skin and the sensor, or variations due to eye blinks. To prevent distortions in the analysis and to ensure validity in the results, we detected and removed such artifacts as follows:

Eye-tracking data. Tobii's default algorithm was used to identify fixations and saccades (for details please see Olsen, 2012). A filter was used to remove the raw gaze points that were classified as blinks.

Facial 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 participant. 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 participant were present, the participant'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, which had the ID value of one, was systematically removed.

Physiological data from a wristband sensor. A simple smoothing function was used to remove any unwanted spikes in the time series in the four data streams coming from Empatica E4 device. This was a simple running average with a moving window of 100 samples and an overlap of 50 samples between two consecutive windows.

Physiological data (such as pupil dilation, HR, BVP, and skin temperature) is susceptible to many personal and contextual biases. These biases include time of day, physical health condition, gender, age, overnight sleep, and so on. All the features, except the facial expression data, were normalized using the first 30 seconds of the data streams to remove the subjective and contextual bias from the data. Thus, for normalization, every data point was expressed as a proportion of the means of the first 30 seconds. Furthermore, the pupil dilation was also normalized with the darkest and the brightest screen shots in the whole

interaction for each participant. Finally, the time series were normalized using the MinMax normalization. None of the participants were removed from the data set.

3.5. Variables

As learner performance depends on internal cognitive processes, we selected debugging performance (i.e., performance) to be our dependent variable. Moreover, expertise is a complex phenomenon that develops with experience over time and is highly contextualized (McCauley et al., 2008). Thus, it was necessary to examine whether expertise affects learner performance and how much explanation power adds to the models. Furthermore, since *we allowed the participants to edit the code and execute it as many times as they want it, rather than finding and reporting bugs, there were frequent switches among the code, errors, and output areas of the IDE*. Consequently, we took these next steps for greater understanding of participants' coding behaviour, as well as for higher validity and reliability of the study results: 1) look into the AOIs fixation proportions, and 2) calculate two-way transitions. The selected variables are continuous, with exception of the variables' use of a debugger and variable view, which are categorical variables.

Expertise. The expertise of the participants was decided by the pre-task test where the students were presented with three small debugging tasks. Each task contained three bugs that did not differ in difficulty. We expected students to remove all bugs within 20 minutes. The level of expertise ranged between 0 and 3, depending on the number of assignments successfully solved by the participants. For example, three bugs per assignment needed to be removed so that the assignment could be counted as solved. Forty percent of participants solved none of the assignments, 25% solved one assignment, 27.5% solved two assignments, and 7.5% solved all three.

Debugging performance. To finish the main debugging task, students were required to solve five questions in a particular order within 40 minutes. After 40 minutes, they were asked to stop, and the number of solved questions, ranging between 0 and 5, was taken as the measure of performance. The percentage of students who solved none of the questions was 17.5; 2.5% solved one question; 17.5% solved two questions; 10% solved three questions; 32.5% solved four questions, and 20% solved all five questions.

Table 2 contains the measures extracted from the collected multimodal data. These measures are informed by the literature (see below), are often used as measures in education and problem solving research (see related work), are ideally suited to capture and measure cognition and affect objectively (Azevedo, 2015), and are of most interest for the purpose of this study.

Eye-tracking related measures. Eye movements are mostly controlled by unconscious processes; hence, they give direct information of learners' perceptual and cognitive processes (Holmqvist et al., 2011). Measured alone, gaze features provide information only about visual processing; however, when combined with learning performance measures, they provide information regarding cognitive processing and mental effort (Mayer, 2010).

Physiological-related measures. Skin conductance, also known as electro-dermal activity (EDA), is an important indicator of nervous system arousal. The features commonly used to define affective states in education in combination with EDA are heart rate (HR), skin temperature, and blood-volume pulse (BVP) (Andreassi, 2010).

Face-related measures. Ekman, Freisen, and Ancoli (1980) developed the Facial Action Coding System to measure facial expressions using "facial actions." During complex learning tasks, students often experience confusion, frustration, boredom, and delight, as reported in previous findings (Craig, Graesser, Sullins, & Gholson, 2004; D'Mello & Graesser, 2012). These are learning-specific or epistemic emotions (McDaniel et al., 2007) and are of interest to our study.

Table 1. Descriptions of the AOIs in Eclipse IDE

Area of Interest (AOI)	Description
Toolbar	The toolbar of the IDE.
Variable View	During a debug, allows changing the value of a variable to test how your program handles a particular value or speeds through a loop.
DebugView	Allows managing the debugging or running a program in the workbench.
ProjectExplorer	Provides a hierarchical view of the artifacts in the workbench.
ProblemView	Automatically log problems, errors, or warnings and groups them by severity.
MainTask	Panel where the code is written.
Question	Five debugging sub-tasks as part of the main method.
Console	Shows the output of the code.

Table 2. Measurements of the Collected Data Streams

Data Stream	Measurement and Definitions
Logs (8 features in total)	**Reading-writing (R-W) episodes:** depicts the difference between the time when students are editing versus the time when they are reading the code. The writing episodes were detected using the frequency of the keystrokes. If there was a gap of more than 40 seconds (i.e., a data-driven threshold) we marked the episode as reading. We compute the mean and standard deviance of R-W episodes (i.e., *4 features*).
	Use of debugger: a binary variable that shows whether students used the debugging support tool from the IDE or not (i.e., *2 features*).
	Use of variable view: a binary variable that shows whether students used the variable declaration and the current value panel from the IDE (i.e., *2 features*).
Face videos (17 features)	**Action Units:** we used OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018) to detect the 17 Action Units (Ekman & Friesen, 1978) from the facial expressions (i.e., a continuous value between 0 and 1 as features for prediction).
Physiological data (7 features)	Empatica E4 provided us with four data streams: heart rate (HR), blood volume pulse (BVP), skin temperature (TMP), and electrodermal activation (EDA). We computed the mean and SD of BVP, TMP, and EDA *(6 features in total)* and the mean of HR *(1 feature)*.
Eye-tracking (40 features in total)	Mean fixation duration *(1 feature)*
	Mean saccade length and duration *(2 features)*
	Pupil diameter *(1 feature)*
	Individual Areas of Interest (AOIs): Eclipse IDE was divided into eight functional AOIs (i.e., *8 features*), which are the basic panels in the interface of the IDE, shown in Figure 1 and described in Table 1. During the analysis, the researchers computed the proportion of time spent on each of the AOIs. The results for the AOIs were later compared to the participant performance and expertise.
	Two-way Transitions among AOIs: We computed a two-way transition probability of the eyes moving from one AOI to any other AOI. This type of transition shows the attention shift from one part of IDE to another, which might correspond to a specific behavioural pattern, or when summed up, a programming strategy. For example, the transition from code to console might depict the behaviour of having a hypothesis about the functionality of the code after changing a few lines of code and checking for the output in the console in order to verify the hypothesis. The sum of all the probabilities of "moving away from or into" one AOI is equal to one. Since there are eight AOIs, we have 28 unique AOI (8*7/2) transitions (i.e., *28 features*).

3.6. Data Analysis

To explore how much explanation MMLA adds on top of the IDE-log data, we created eight predictive models (Table 3). We define the model based on the data and the classifier. Thus, once the data changes, the model changes itself. For example, model 1 (M1) is defined by the measures from the IDE-log data and a Random Forest (RF) classifier, while M2 is defined by the measures from the IDE-log data, gaze data, and a RF classifier. M1 is the basic model because it includes only measures from IDE-log data. The basic model was used to compare the value with the other models (i.e., M2–M8), which are augmented with features from multimodal data. The data was segmented using a variable length time window. Each time window corresponds to one student solving one debugging problem. Table 3 summarizes the scheme that we used to build the predictive models.

Table 3. Predictive Models Built with Multimodal Data

Model	Dependent Variable	Independent Variable
M1	Debugging performance	Expertise + Features from log data [Reading episode + Writing episode + Debugger View + Variable View + N° of times the main method was run + Time difference between two main method runs]
M2	Debugging performance	M1 + features from eye-tracking data
M3	Debugging performance	M1 + features from facial data
M4	Debugging performance	M1 + features from physiological data*
M5	Debugging performance	M1 + features from eye-tracking and facial data
M6	Debugging performance	M1 + features from eye-tracking and physiological data*
M7	Debugging performance	M1 + features from facial and physiological data*
M8	Debugging performance	M1 + features from eye-tracking, facial, and physiological data*

* Physiological data from wristband sensor.

For our analysis we decided to use RF as the predictive algorithm. Since we have continuous and categorical variables, combining them in one classifier other than RF was difficult. Moreover, RF offers easy extraction of features importance, and has been found to be a top-performing algorithm in a large comparative study (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). We also looked into neural networks (NNs), but since the outcomes cannot be explainable in terms of the way the NN algorithm combines the features to produce the output, and we cannot calculate the importance of features, we concluded that NNs were not a smart approach. Finally, we looked into the support vector machine (SVM), but since SVM can be biased based on the chosen support vectors, we decided to avoid the risk of overfitting the models.

Next, we divided the whole dataset into training (90% of the students) and testing (10% of the students) sets. To remove the selection bias in the training set, we performed a 10-fold cross validation. All models used the following scheme of training-validation-testing. First, a 10% sample of the data (based on student ID) was kept aside for "out of sample" testing. Then, the remaining 90% of the data was used to train the model using a 10-fold cross-validation. In each cross-validation step, 10% of the data was used to validate the training. The best model from the 10-fold cross-validation was finally tested on the data kept for "out of sample testing." This reduced the variability and presented more accurate estimates of performance. We also employed precision, recall, and f-measure, to evaluate the robustness of the models. Finally, in order to compare the models, we used adjusted R^2 as an evaluation metric that considers the number of estimated parameters while evaluating the models. The adjusted R^2 does not increase if the number of predictors increases; it increases only if the new predictor improves the model more than what would be expected by chance.

4. Results

The models from M2 to M8 significantly outperformed M1 (see Figure 3), with M1 exhibiting a significantly lower adjusted R^2 value of 0.42 (std.dev. = 0.116). The biggest improvement appears in M8 (R^2= 0.81; std.dev. = 0.142) and the smallest in M4 (R^2= 0.54; std.dev. = 0.102). The eight models are presented in Figure 3 with the adjusted R^2 values and their 95% confidence interval. In addition, Table 4 depicts the statistical difference between the R^2 values of the eight models by conducting a t-test, identifying differences that in some cases are relatively strong.
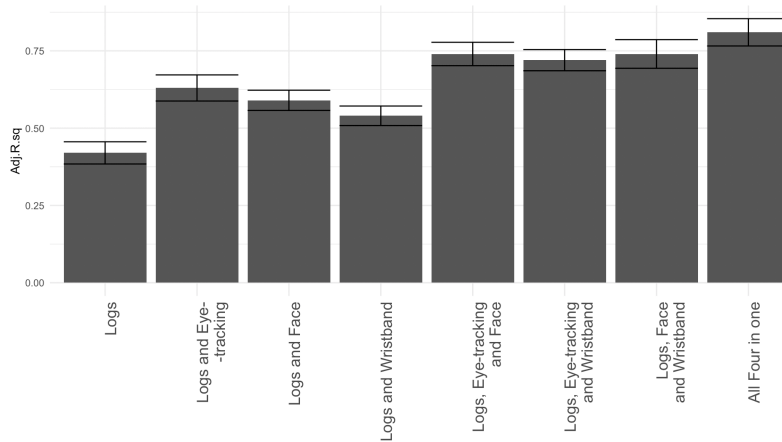
Figure 3. Adjusted R^2 values and the confidence interval (95%) of the eight prediction models

Table 4. Pairwise Model Comparison

	M2	M3	M4	M5	M6	M7	M8
M1	**3.89****	**3.61****	**2.55***	**6.21*****	**6.09*****	**5.68*****	**6.92*****
M2	-	0.77	1.73	1.96	1.67	1.82	**2.97***
M3	-	-	1.11	**3.03****	**2.76***	**2.75***	**4.04****
M4	-	-	-	**4.04****	**3.82****	**3.67****	**4.96****
M5	-	-	-	-	0.38	0.00	1.20
M6	-	-	-	-	-	0.32	1.43
M7	-	-	-	-	-	-	1.12

* p < .05, ** p < .01, *** p < .001

To evaluate the robustness of the eight models, we also computed the precision, recall, and f-measure based on majority class prediction baseline. The results are depicted in Table 5, showcasing relatively high values.

Table 5. Precision, Recall, and F-Measure

	precision	recall	F-measure
M1	0.73	0.68	0.70
M2	0.82	0.80	0.81
M3	0.80	0.76	0.77
M4	0.78	0.74	0.76
M5	0.85	0.83	0.84
M6	0.84	0.81	0.82
M7	0.83	0.80	0.81
M8	0.86	0.84	0.85

Next, we performed features importance analysis; that is, we calculated the importance with regards to predicting student performance of the extracted individual measures/features (Table 2) for each data stream. As shown in Table 2, from the ***log data*** we extracted eight features, ranked by importance in Figure 4. One can observe that the expertise is the most important feature in predicting performance, with the number of times the main method was run, and the time difference between two runs, as the second and third most significant features.

From the ***eye-tracking data,*** we extracted 40 features (Table 2), with Figure 5 depicting the 12 most important features (on a scale from 100 to zero). The most important feature for predicting performance is the mean saccade length; next in line is the pupil diameter, which in research is considered a physiological marker of processing load, cognitive effort, or use of attentional resources (Buettner, 2013).

From the ***facial data*** we extracted 17 features (i.e., action units), all ranked in importance on a scale from 100 to zero (Figure 6). In our study, the most important features for predicting performance are nose wrinkle, followed by blinking, both coming from the upper part of the face.
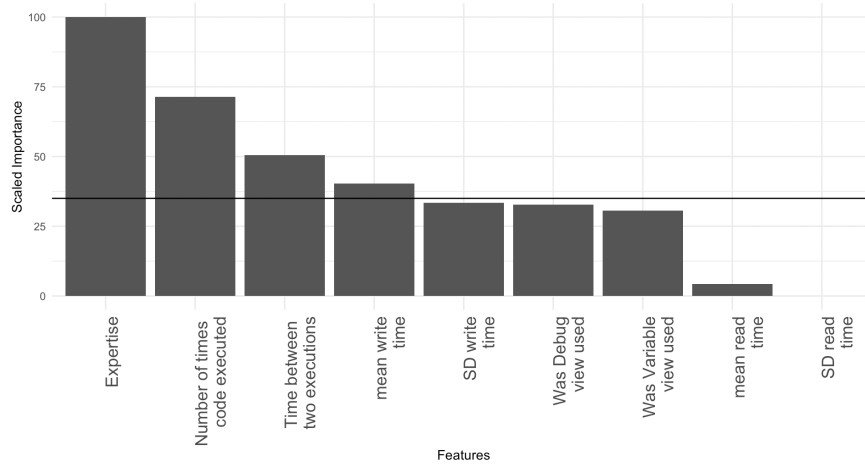
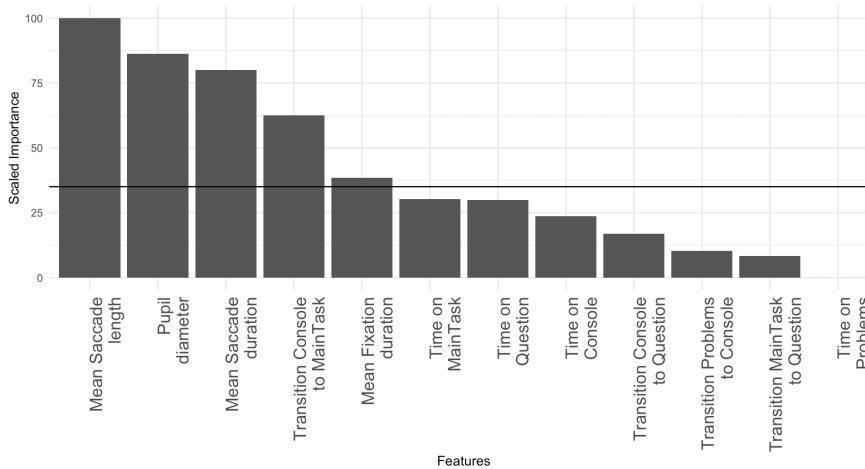**Figure 4.** Feature importance for log data scaled from 100 to zero

Figure 5. Feature importance for eye-tracking data scaled from 100 to zero
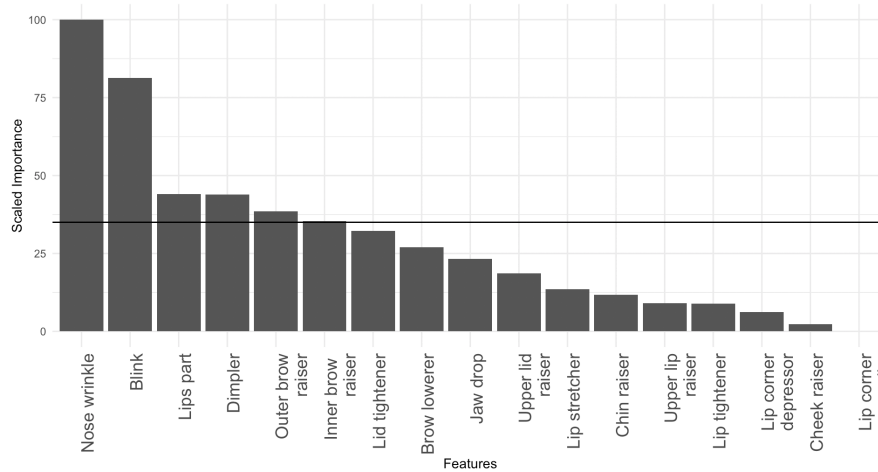

Figure 6. Feature importance for facial data scaled from 100 to zero

Finally, from the ***physiological data*** captured by the wristband sensor, we extracted seven features (Table 2). From all seven features (Figure 7), the stress measured by skin temperature (i.e., average temperature of the skin) is the most important predictor of performance, followed by arousal, measured by the BVP variability.
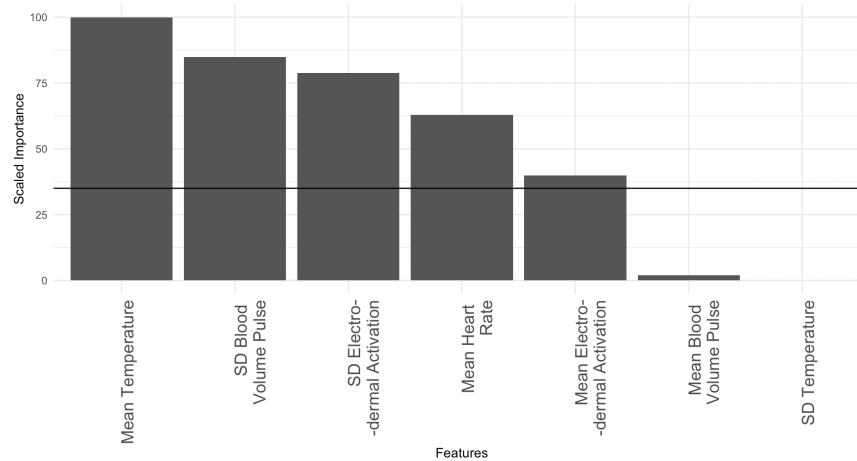
Figure 7. Feature importance for physiological data scaled from 100 to zero

5. Discussion and conclusion

5.1. Interpreting the Results with Respect to the Research Questions

As educators continue to struggle to understand what technology to use and how it can improve learning (Perrotta, 2013), MMLA offers an additional lens through which to investigate student experiences when learning with technology (Blikstein, 2013; Cukurova, Giannakos, & Martinez-Maldonado, 2020; Ochoa, 2017; Oviatt et al., 2018). Findings generated with MMLA can inform the (re)design of learning activities and objectives in computer-mediated learning settings (Järvelä, Malmberg, Haataja, Sobocinski, & Kirschner, 2019; Mangaroska & Giannakos, 2018), because they capture learner behaviour and actions in real time and in an objective way, along the cognitive and affective dimensions of learning (Azevedo, 2015).

To elaborate on *the potential of MMLA in augmenting traditional LA with respect to LD in the context of CE (RQ1)* we refer to Table 4. Hence, adding measures from individual multimodal streams of data (e.g., gaze data, facial data, or physiological data) on top of the IDE-log data, as shown through M2, M3, and M4, demonstrated improvement compared to the basic M1 model, but less improvement compared to M5, M6, M7, and M8, where we had combinations of measures from more than one data stream. These improvements are evident despite the fact that *expertise (i.e., previous knowledge)* was included in the basic model. In other words, our findings do not ignore the fact that expertise is a valuable indicator of learner performance in programming and problem solving, but also indicate the value of MMLA regardless of student experience. Overall, our results support the premise that *behaviour captured by multimodal data provides additional understanding of learner performance (e.g., specific behavioural patterns when moving from one AOI to another during hypothesis verification) that cannot be captured with programming process data as expertise can*. Moreover, the model with gaze data (i.e., M2) is improved only by the model that carries all modalities (i.e., M8). This is in line with previous work that demonstrated *gaze data to be a key element in understanding and predicting learner behaviour and performance in problem solving activities due to the direct relation with cognitive load and mental effort* (Beatty, 1982; Bednarik, 2012; Bednarik et al., 2020; Kaller, Rahm, Bolkenius, & Unterrainer, 2009).

On the other hand, facial and physiological data combined on top of M2 (i.e., M7), provided a significantly higher prediction rate than M2. This finding suggests **that** *facial and physiological data provide different information than gaze data (i.e., M2).* Because of that, researchers should not be ignorant in using facial and physiological data, both in terms of *creating effective LDs,* as well *as informing measures and models used in LA* (D'Mello & Graesser, 2012; D'Mello, Lehman, Pekrun, & Graesser, 2014)*.* Furthermore, combining features from facial, physiological, and log data, created M7, which is easily comparable to the best model in our analysis (i.e., M8). From a design perspective, cameras and wristband sensors carry the notion of "at scale" (Ogan, 2019). Therefore, through the prism of LD, the ubiquitous nature of facial and physiological data can have *strong design potential at a classroom level* (Ogan, 2019), *where measures from models build on these data streams can easily be used to* **inform LD decisions for classroom orchestration** (Dillenbourg, 2013).

To strengthen the link and move forward the alignment between LA and LD, we need a consolidated model of theory, design, and data science, as proposed by Gašević et al. (2017). So far analyzing LDs considering mostly the pedagogical intent, has supported the creation of valuable models and frameworks of LA (Mangaroska & Giannakos, 2018). However, these models and frameworks are insufficiently linked with foundational disciplines, in particular the LS (Gašević et al., 2017; Neelen & Kirschner, 2017; Reimann, 2016; Wiley, Dimitriadis, Bradford, & Linn, 2020). This creates a lack of *evidence-*

informed practices that can help educators to understand how and when their students learn best, so that the planning, execution, and evaluation of their LDs (including the learning objectives) can be more effective (Kirschner, 2017). Thus, if we aim for LA that are consequential for learning, and can support educators to create effective LDs and optimize the process of learning, we need to attach value to the evidence from the LS and synthesize it with design and data science, as a common comprehensive language for educators to use and act upon in their own educational practice (Gašević et al., 2017).

To answer the second RQ and draw implications for potential further advancement of the synergy between LA and LD in computing education, we performed a feature importance analysis. The most significant features from **log data** (Figure 4) indicate the presence of strategies similar to the ones found by Sharma et al. (2018) and Turkle and Papert (1992). The ***number of main method runs*** and ***the time difference between two runs*** are measures that educators can use to model learning strategies in programming (e.g., "thinkers" vs. "probers"), as there is already evidence for the significance of these measures in the literature. Moreover, the time difference between the two runs might also be influenced by the average duration of the writing episodes, signalling that some students might be slow at writing and not necessarily make less change in the code size. These examples show the potential ***capacity that log data has to inform on learning strategies*** and ***facilitate the design of personalized interventions.***

The most significant features from **gaze data** (Figure 5) are considered to be physiological markers of cognitive load, mental effort, and sustained attention (Hollender, Hofmann, Deneke, & Schmitz, 2010; May, Kennedy, Williams, Dunlap, & Brannan, 1990). According to CLT, an increase in mental effort is desirable in complex learning tasks that facilitate deep learning (Hollender et al., 2010). Thus, measures such as ***pupil diameter, average fixation duration,*** and ***average saccade duration***, can assist educators to know when learning activities cause an undesirable level of mental effort that can transition into cognitive "overload," and adjust the LD accordingly. Furthermore, ***measures from gaze transitions*** can be used for developing LA feedback tools that can orchestrate basic behavioural regulation (Mangaroska et al., 2018), such as visual attention or following instructions (Baumeister & Vohs, 2004) when students enter into educationally unproductive states.

The most significant features from **facial data** (Figure 6) and ***physiological data*** (Figure 7) must be combined or linked with specific actions during learning to make sense of and unfold the complex coordination of cognitive-affective states. Looking at action units individually makes it difficult to make the link between an action unit (e.g., lips apart) and an action during coding (e.g., syntax error; McDaniel et al., 2007). Even in face-to-face non-verbal communication, humans tend to connect more features from the face to understand a facial expression (Ekman, 1993). Considering the mapping that McDaniel et al. (2007) provided, we can link specific affective states of learners to particular learning design instructions. Such mapping can aid educators in ***supervising and managing undesirable cognitive-affective learner states*** (e.g., frustration, boredom) in digital settings, ***to mitigate disengagement*** and ***prevent negative cognitive loops and partial learning*** (Craig et al., 2004; D'Mello et al., 2014).

Using machine learning, we looked at the overall patterns in the data and performed feature importance among the 72 measures extracted from the multimodal data. Findings like ours combined with pedagogical intent from educators and theories from the LS, can advance the synergy between LA and LD by translating results in applicable design guidelines that can lead to improvements in the design of learning activities, instructional methods for teaching particular skills, and even the overall course (re)design. The complexity of the MMLA approach is congruous with learning theories because it can be used to understand how effectively students use the opportunities for learning as given in the LD. Such understanding promises to support ***versatile improvements in the LD in digital environments***, from setting the right feedback loop (e.g., explaining misconceptions vs. challenging the student), to the design of personalized interventions, and modelling effective learning strategies considering skills and knowledge proficiency.

5.2. Implications

Our paper aims to provoke deeper, more productive reflections and conversations about the potential of MMLA to expand and advance the synergy of LA and LD among the community of educators from a post-evaluation design-aware process to permanent monitoring process of adaptation (Schmitz, Van Limbeek, Greller, Sloep, & Drachsler, 2017). For now, we offer the following implications that should further inform the research and practice of LA and LD:

- Pedagogical intent alone (Bakharia et al., 2016) is not sufficient for expanding learning analytics that have the capacity to align with learning design and learning theories from the learning sciences; thus, exploration of the overall patterns in the data should also be considered, as it can identify and validate meaningful associations (Rienties et al., 2017) for which educators might be ignorant.

- Different data streams provide different information for learner behaviour, and offer different design potential. For

example, facial and physiological data have strong design potential at the classroom level (Ogan, 2019), while gaze data has design potential at the individual level (Bednarik et al., 2020).

- To optimize learning and make accurate changes in the LD, we need to understand how and when students learn best. Using mainstream LA, without considering the context is not always sufficient. For that we need understanding of learner states (i.e., cognitive, motivational, emotional) during learning, and knowledge of evidence-informed practices from the LS (Kirschner, 2017; Nathan & Wagner Alibali, 2010) to compare with and make adjustments/improvements accordingly.
- Relevant LA (theory informed and grounded in evidence) can inform on domain-specific strategies (e.g., what steps an individual uses to reach a solution) and help to understand how effectively the opportunities for learning as given in the LD are used (Mangaroska et al., 2018).
- MMLA research in CE should extend the taxonomy of design dimensions for IDE-based interventions proposed by Hundhausen et al. (2017), and set the foundation for developing consolidated model of theory, design, and data science in CE. Moreover, along with the increase in MMLA studies, datasets should be made publicly available, so that researchers can compare, consolidate, and verify theories and results.

5.3. Limitations

Our findings speak in favour of the potential of MMLA, but are subjected to certain limitations. Although MMLA methods for measuring complex internal conditions are objective and indirect, as defined by Brunken, Plass, and Leutner (2003), we do not exclude the possibility of measuring complex internal conditions with other methods such as think-aloud protocols, pre-tests/post-tests, or self-reporting questionnaires (Azevedo, 2015). This might shed light on outcomes different than the ones reported in this study, but those outcomes would be based on recalled behaviour and subjective impressions, rather than on learners' actual behaviour. Moreover, our findings originate from authentic learning tasks based on problem-solving practised by an individual; thus, the generated features might be characteristic for individual learning activities where students engage with a computer screen for the entire activity, compared to learning activities where students work collaboratively and with less focus on computer devices. Next, the study was performed in a controlled environment that might affect its ecology because participants signed a consent form and were aware of the physiological data collection, which may increase the desire "to perform" and generate good biomarker metrics. Furthermore, in our analysis we applied only one ML algorithm (i.e., Random Forest) and have not explored what outcomes we might get by applying different or combined algorithms on different data streams. However, we consider the chosen approach to be appropriate and relevant to establish examples of how MMLA can advance the synergy between LA and LD. Finally, the LD guidelines based on features importance analysis are only explored in an initial way and have not been applied to make changes in the current LD; thus, we welcome future studies to further explore the implementation of changes based on the reported measures and recommendations.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

References

Ahonen, L., Cowley, B., Torniainen, J., Ukkonen, A., Vihavainen, A., & Puolamäki, K. (2016). Cognitive collaboration found in cardiac physiology: Study in classroom environment. *PloS One, 11*(7), e0159178. https://dx.doi.org/10.1371/journal.pone.0159178

Altadmri, A., & Brown, N. C. (2015). 37 million compilations: Investigating novice programming mistakes in large-scale student data. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (SIGCSE '15), 4–7 March 2015, Kansas City, MO, USA (pp. 522–527). New York, NY: ACM. https://dx.doi.org/10.1145/2676723.2677258

Andreassi, J. L. (2010). *Psychophysiology: Human behavior and physiological response*. Hove, East Sussex, UK: Psychology Press.

Arapakis, I., Lalmas, M., & Valkanas, G. (2014). Understanding within-content engagement through pattern analysis of mouse gestures. *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management* (CIKM '14), 3–7 November 2014, Shanghai, China (pp. 1439–1448). New York, NY: ACM. https://dx.doi.org/10.1145/2661829.2661909

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

Baker, R. S., 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://dx.doi.org/10.1016/j.ijhcs.2009.12.003

Bakharia, A., Corrin, L., De Barba, P., Kennedy, G., Gašević, D., Mulder, R., … Lockyer, L. (2016). A conceptual framework linking learning design with learning analytics. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 329–338). New York: ACM. https://dx.doi.org/10.1145/2883851.2883944

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition* (FG 2018), 15–19 May 2018, Xi'an, China (pp. 59–66). Washington, DC: IEEE Computer Society. https://dx.doi.org/10.1109/FG.2018.00019

Baumeister, R. F., & Vohs, K. D. (2004). *Handbook of self-regulation: Research, theory, and applications*. New York, NY: The Guilford Press.

Beattie, V., Collins, B., & McInnes, B. (1997). Deep and surface learning: A simple or simplistic dichotomy? *Accounting Education, 6*(1), 1–12. https://dx.doi.org/10.1080/096392897331587

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276. https://dx.doi.org/10.1037/0033-2909.91.2.276

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://dx.doi.org/10.1016/j.ijhcs.2011.09.003

Bednarik, R., Busjahn, T., Gibaldi, A., Ahadi, A., Bielikova, M., Crosby, M., … Lister, R. (2020). EMIP: The eye movements in programming dataset. *Science of Computer Programming*, *198*, article 102520. https://dx.doi.org/10.1016/j.scico.2020.102520

Begel, A. (2016). *Fun with software developers and biometrics: Invited talk. Proceedings of the 1st International Workshop on Emotion Awareness in Software Engineering* (SEmotion '16), 17 May 2016, Austin, TX, USA (pp. 1–2). New York, NY: ACM. https://dx.doi.org/10.1145/2897000.2897007

Blikstein, P. (2013). Multimodal learning analytics. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 102–106). New York, NY: ACM. https://dx.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://dx.doi.org/10.18608/jla.2016.32.11

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of the Learning Sciences, 23*(4), 561–599. https://dx.doi.org/10.1080/10508406.2014.954750

Bruce, C., Buckingham, L., Hynd, J., McMahon, C., Roggenkamp, M., & Stoodley, I. (2004). Ways of experiencing the act of learning to program: A phenomenographic study of introductory programming students at university. *Journal of Information Technology Education: Research, 3*(1), 145–160. https://dx.doi.org/10.28945/294

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53–61. https://dx.doi.org/10.1207/S15326985EP3801_7

Buettner, R. (2013). *Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology.* Annual Conference on Artificial Intelligence (pp. 37–48). Lecture notes in Computer Science, vol. 8077. https://dx.doi.org/10.1007/978-3-642-40942-4_4

Busjahn, T., Schulte, C., Sharif, B., Begel, A., Hansen, M., Bednarik, R., … Antropova, M. (2014). Eye tracking in computing education. *Proceedings of the 10th Annual Conference on International Computing Education Research* (ICER '14), 11–13 August 2014, Glasgow, UK (pp. 3–10). New York, NY: ACM. https://dx.doi.org/10.1145/2632320.2632344

Carter, A. S., & Hundhausen, C. D. (2015). The design of a programming environment to support greater social awareness and participation in early computing courses. *Journal of Computing Sciences in Colleges, 31*(1), 143–153.

Chan, M. C. E., Ochoa, X., & Clarke, D. (2020). Multimodal learning analytics in a laboratory classroom *Machine Learning Paradigms* (pp. 131–156): Springer. https://dx.doi.org/10.1007/978-3-030-13743-4_8

Corrin, L., De Barba, P., Lockyear, L., Gašević, D., Williams, D., Dawson, S., … Bakharia, A. (2016). *Completing the loop: Returning meaningful learning analytic data to teachers*. Retrieved from https://research.monash.edu/en/publications/completing-the-loop-returning-meaningful-learning-analytic-data-t

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://dx.doi.org/10.1080/1358165042000283101

Crk, I., & Kluthe, T. (2014). Toward using alpha and theta brain waves to quantify programmer expertise. *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC 2014), 26–30 August 2014, Chicago, IL, USA. Washington, DC: IEEE Computer Society. https://dx.doi.org/10.1109/embc.2014.6944840

Cukurova, M., Giannakos, M., & Martinez-Maldonado, R. (2020). The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology, 51*(5), 1441–1449. https://dx.doi.org/10.1111/bjet.13015

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://dx.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://dx.doi.org/10.1016/j.learninstruc.2012.05.003

Dillenbourg, P. (2013). Design for classroom orchestration. *Computers & Education, 69*, 485–492. https://dx.doi.org/10.1016/j.compedu.2013.04.013

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*, 1153. https://dx.doi.org/10.3389/fpsyg.2017.01153

Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 4–9 May 2019, Glasgow, Scotland, UK (Paper No. 39). New York, NY: ACM. https://dx.doi.org/10.1145/3290605.3300269

Edwards, S. H., & Perez-Quinones, M. A. (2008). Web-CAT: Automatically grading programming assignments. *ACM SIGCSE Bulletin, 40*(3), 328. https://dx.doi.org/10.1145/1384271.1384371

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://dx.doi.org/10.1037/0003-066X.48.4.384

Ekman, P., & Friesen, E. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Freisen, W. V., & Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology, 39*(6), 1125-1134. https://dx.doi.org/10.1037/h0077722

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Fritz, T., Begel, A., Müller, S. C., Yigit-Elliott, S., & Züger, M. (2014). Using psycho-physiological measures to assess task difficulty in software development. *Proceedings of the 36th International Conference on Software Engineering* (ICSE 2014), 31 May–7 June 2014, Hyderabad, India (pp. 402–413). New York, NY: ACM. https://dx.doi.org/10.1145/2568225.2568266

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

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice, 3*(1), 63–78. https://dx.doi.org/10.1080/23735082.2017.1286142

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management, 48*, 108–119. https://dx.doi.org/10.1016/j.ijinfomgt.2019.02.003

Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., & Divakaran, A. (2016). Multimodal analytics to study collaborative problem solving in pair programming. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 516–517). New York: ACM. https://dx.doi.org/10.1145/2883851.2883877

Hernández-Leo, D., Martinez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J. A., & Rodríguez-Triana, M. J. (2019). Analytics for learning design: A layered framework and tools. *British Journal of Educational Technology, 50*(1), 139–152. https://dx.doi.org/10.1111/bjet.12645

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://dx.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, UK: Oxford University Press.

Hundhausen, C. D., Olivares, D. M., & Carter, A. S. (2017). IDE-based learning analytics for computing education: A process model, critical review, and research agenda. *ACM Transactions on Computing Education (TOCE), 17*(3), 11, 1–26. https://dx.doi.org/10.1145/3105759

Ihantola, P., Sorva, J., & Vihavainen, A. (2014). Automatically detectable indicators of programming assignment difficulty. *Proceedings of the 15th Annual Conference on Information Technology Education* (SIGITE '14), 15–18 October 2014, Atlanta, GA, USA (pp. 33–38). New York, NY: ACM. https://dx.doi.org/10.1145/2656450.2656476

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., … Rivers, K. (2015). Educational data mining and learning analytics in programming: Literature review and case studies. *Proceedings of the 2015 ITiCSE on Working Group Reports* (ITICSE-WGR '15), 4–8 July 2015, Vilnius, Lithuania (pp. 41–63). New York, NY: ACM. https://dx.doi.org/10.1145/2858796.2858798

Jadud, M. C. (2006). Methods and tools for exploring novice compilation behaviour. *Proceedings of the 2nd International Workshop on Computing Education Research* (ICER '06), 9–10 September 2006, Canterbury, UK (pp. 73–84). New York, NY: ACM. https://dx.doi.org/10.1145/1151588.1151600

Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2019). What multimodal data can tell us about the students' regulation of their learning process. *Learning and Instruction* (in press, corrected proof). https://dx.doi.org/10.1016/j.learninstruc.2019.04.004

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://dx.doi.org/10.1111/j.1469-8986.2009.00821.x

Kevic, K., Walters, B. M., Shaffer, T. R., Sharif, B., Shepherd, D. C., & Fritz, T. (2015). Tracing software developers' eyes and interactions for change tasks. *Proceedings of the 10th Joint Meeting on Foundations of Software Engineering* (ESEC/FSE 2015), 30 August–4 September 2015, Bergamo, Italy (pp. 202–213). New York, NY: ACM. https://dx.doi.org/10.1145/2786805.2786864

Kirschner, P. A. (2017). *If it ain't there, it's broke!* Retrieved from https://3starlearningexperiences.wordpress.com/2017/09/05/if-it-aint-there-its-broken/

Kress, G. (2001). *Multimodal teaching and learning: The rhetorics of the science classroom*. London, UK: A&C Black.

Kurland, D. M., Pea, R. D., Clement, C., & Mawby, R. (1986). A study of the development of programming ability and thinking skills in high school students. *Journal of Educational Computing Research, 2*(4), 429–458. https://dx.doi.org/10.2190%2FBKML-B1QV-KDN4-8ULH

Lockyer, L., & Dawson, S. (2012). Where learning analytics meets learning design. In S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 14–15). New York, NY: ACM. https://dx.doi.org/10.1145/2330601.2330609

Mangaroska, K., & Giannakos, M. N. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, *12*(4), 516–534. https://dx.doi.org/ 10.1109/TLT.2018.2868673

Mangaroska, K., Sharma, K., Giannakos, M., Træteberg, H., & Dillenbourg, P. (2018). Gaze-driven design insights to amplify debugging skills: A learner-centred analysis approach. *Journal of Learning Analytics, 5*(3), 98–119. https://dx.doi.org/10.18608/jla.2018.53.7

Marion, B., Impagliazzo, J., St. Clair, C., Soroka, B., & Whitfield, D. (2007). Assessing computer science programs: What have we learned? *Proceedings of the 38th ACM Technical Symposium on Computer Science Education* (SIGCSE 07), 7–10 March 2007, Covington, KY, USA (pp. 131–132). New York, NY: ACM. https://dx.doi.org/10.1145/1227310.1227358

Martinez-Maldonado, R., Mangaroska, K., Schulte, J., Elliott, D., Axisa, C., & Shum, S. B. (2020). 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), Article No. 22. https://dx.doi.org/10.1145/3381017

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica, 75*(1), 75–89. https://dx.doi.org/10.1016/0001-6918(90)90067-P

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*(2), 167–171. https://dx.doi.org/10.1016/j.learninstruc.2009.02.012

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://dx.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 D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (CogSci 2007), 1–4 August 2007, Nashville, TN, USA (pp. 467–472). Austin, TX: Cognitive Science Society. http://csjarchive.cogsci.rpi.edu/Proceedings/2007/docs/p467.pdf

Melero, J., Hernández-Leo, D., Sun, J., Santos, P., & Blat, J. (2015). How was the activity? A visualization support for a case of location-based learning design. *British Journal of Educational Technology, 46*(2), 317–329. https://dx.doi.org/10.1111/bjet.12238

Müller, S. C. (2015). Measuring software developers' perceived difficulty with biometric sensors. *Proceedings of the 37th International Conference on Software Engineering* (ICSE '15), 16–24 May 2015, Florence, Italy (Vol. 2, pp. 887–890). New York, NY: ACM. https://dx.doi.org/ 10.1109/ICSE.2015.284

Nathan, M. J., & Wagner Alibali, M. (2010). Learning sciences. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(3), 329–345. https://dx.doi.org/10.1002/wcs.54

Neelen, M., & Kirschner, P. (2017). *Where are the learning sciences in learning analytics research?* Retrieved from https://3starlearningexperiences.wordpress.com/2017/10/17/where-are-the-learning-sciences-in-learning-analytics-research/

O'Grady, M. J. (2012). Practical problem-based learning in computing education. *ACM Transactions on Computing Education (TOCE), 12*(3), 1–16. https://dx.doi.org/10.1145/2275597.2275599

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. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 360–364). New York, NY: ACM. https://dx.doi.org/10.1145/3170358.3170406

Ochoa, X. (2017). Multimodal learning analytics. In C. Lang, G. Siemens, A. Wise & D. Gašević (Eds.), *The Handbook of Learning Analytics* (1st Ed.) (pp. 129–141). Society for Learning Analytics Research. https://dx.doi.org/10.18608/hla17

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics, 3*(2), 213–219. https://dx.doi.org/10.18608/jla.2016.32.10

Ogan, A. (2019). Reframing classroom sensing: Promise and peril. *Interactions, 26*(6), 26–32. Retrieved from https://interactions.acm.org/archive/view/november-december-2019/reframing-classroom-sensing

Olsen, A. (2012, March 20). The Tobii I-VT fixation filter: Algorythm description. *Tobii Technology*. Retrieved from https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf

Oviatt, S., Grafsgaard, J., Chen, L., & Ochoa, X. (2018, October). Multimodal learning analytics: Assessing learners' mental state during the process of learning. In S. Oviatt, B. Schuller & P. R. Cohen (Eds.), *The handbook of multimodal-multisensor interfaces: Signal processing, architectures, and detection of emotion and cognition* (Vol. 2) (pp. 331–374). New York, NY: ACM Books. https://dx.doi.org/10.1145/3107990.3108003

Parnin, C. (2011). Subvocalization: Toward hearing the inner thoughts of developers. *Proceedings of the IEEE 19th International Conference on Program Comprehension* (ICPC 2011) 22–24 June 2011, Kingston, ON, Canada (pp. 197–200). New York, NY: ACM. https://dx.doi.org/10.1109/ICPC.2011.49

Perkins, D. N., Hancock, C., Hobbs, R., Martin, F., & Simmons, R. (1986). Conditions of learning in novice programmers. *Journal of Educational Computing Research, 2*(1), 37–55. https://dx.doi.org/10.2190%2FGUJT-JCBJ-Q6QU-Q9PL

Perrotta, C. (2013). Do school-level factors influence the educational benefits of digital technology? A critical analysis of teachers' perceptions. *British Journal of Educational Technology, 44*(2), 314–327. https://dx.doi.org/10.1111/j.1467-8535.2012.01304.x

Persico, D., & Pozzi, F. (2015). Informing learning design with learning analytics to improve teacher inquiry. *British Journal of Educational Technology, 46*(2), 230–248. https://dx.doi.org/10.1111/bjet.12207

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://dx.doi.org/10.1111/jcal.12232

Reimann, P. (2016). Connecting learning analytics with learning research: The role of design-based research. *Learning: Research and Practice, 2*(2), 130–142. https://dx.doi.org/10.1080/23735082.2016.1210198

Rienties, B., Nguyen, Q., Holmes, W., & Reedy, K. (2017). A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK. *Interaction Design and Architecture(s), 33*, 134–154.

Rodríguez-Triana, M. J., Prieto, L. P., Martínez-Monés, A., Asensio-Pérez, J. I., & Dimitriadis, Y. (2018). The teacher in the loop: Customizing multimodal learning analytics for blended learning. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 417–426). New York, NY: ACM. https://dx.doi.org/10.1145/3170358.3170364

Schmitz, M., Van Limbeek, E., Greller, W., Sloep, P., & Drachsler, H. (2017). Opportunities and challenges in using learning analytics in learning design. *Proceedings of the 12th European Conference on Technology Enhanced Learning* (EC-TEL 2017), 12–15 September 2017, Tallinn, Estonia (pp. 209–223). Lecture Notes in Computer Science, vol. 10474. Springer. https://dx.doi.org/10.1007/978-3-319-66610-5_16

Schulte, C., Magenheim, J., Müller, K., & Budde, L. (2017). The design and exploration cycle as research and development framework in computing education. *Proceedings of the 2017 IEEE Global Engineering Education Conference* (EDUCON 2017), 25–28 April 2017, Athens, Greece (pp. 867–876). Washington, DC: IEEE Computer Society. https://dx.doi.org/10.1109/EDUCON.2017.7942950

Sharma, K., Mangaroska, K., Trætteberg, H., Lee-Cultura, S., & Giannakos, M. (2018). Evidence for programming strategies in university coding exercises. *Proceedings of the 13th European Conference on Technology Enhanced Learning* (EC-TEL 2018)*, 3–5 September 2018, Leeds, UK (pp. 326–339). Lecture Notes in Computer Science, vol. 11082. Cham, Switzerland: Springer. https://dx.doi.org/10.1007/978-3-319-98572-5_25

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://dx.doi.org/10.3102%2F0013189X015002004

Snow, R. E., Corno, L., & Jackson III, D. N. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243-310). New York, NY: Macmillan Library Reference USA.

Turkle, S., & Papert, S. (1992). Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior, 11*(1), 3–33.

Van Harmelen, M., & Workman, D. (2012). Analytics for learning and teaching. *CETIS Analytics Series, 1*(3), 1–40.

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5–13. https://dx.doi.org/10.1207/S15326985EP3801_2

Wiley, K. J., Dimitriadis, Y., Bradford, A., & Linn, M. C. (2020). From theory to action: Developing and evaluating learning analytics for learning design. *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK '20), 23–27 March 2020, Frankfurt, Germany (pp. 569–578). New York, NY: ACM. https://dx.doi.org/10.1145/3375462.3375540

Wise, A. F., & Jung, Y. (2019). Teaching with Analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics, 6*(2), 53–69. https://dx.doi.org/10.18608/jla.2019.62.4

Worsley, M. (2014). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 1–4). New York, NY: ACM. https://dx.doi.org/10.1145/2666633.2666634

Worsley, M., & Blikstein, P. (2013). Towards the development of multimodal action based assessment. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 94–101). New York: ACM. https://dx.doi.org/10.1145/2460296.2460315

Worsley, M., & Blikstein, P. (2015). Leveraging multimodal learning analytics to differentiate student learning strategies. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 360–367). New York, NY: ACM. https://dx.doi.org/10.1145/2723576.2723624