Author(s): Lämsä, Joni; Mannonen, Joonas; Tuhkala, Ari; Heilala, Ville; Helovuo, Arto; Tynkkynen, Ilkka; Lampi, Emilia; Sipiläinen, Katriina; Kärkkäinen, Tommi; Hämäläinen, Raija

Title: Capturing cognitive load management during authentic virtual reality flight training with behavioural and physiological indicators

Year: 2023

Version: Published version

Received: 29 November 2021 | Revised: 10 March 2023 | Accepted: 10 April 2023

DOI: 10.1111/jcal.12817

ORIGINAL ARTICLE

Journal of Computer Assisted Learning WILEY

Capturing cognitive load management during authentic virtual reality flight training with behavioural and physiological indicators

Joni Lämsä[1,2] | Joonas Mannonen[3] | Ari Tuhkala[1] | Ville Heilala[4] | Arto Helovuo[5] | Ilkka Tynkkynen[5] | Emilia Lampi[3] | Katriina Sipiläinen[6] | Tommi Kärkkäinen[7] | Raija Hämäläinen[1]

[1]Department of Education, University of Jyväskylä, Jyväskylä, Finland

[2]Learning and Educational Technology (LET) Research Lab, University of Oulu, Oulu, Finland

[3]Finnish Institute for Educational Research, University of Jyväskylä, Jyväskylä, Finland

[4]Faculty of Information Technology, University of Jyväskylä; Department of Education, University of Jyväskylä, Jyväskylä, Finland

[5]Finnair, Helsinki-Uusimaa, Finland

[6]Faculty of Education and Psychology, University of Jyväskylä, Jyväskylä, Finland

[7]Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

Correspondence
Joni Lämsä, Learning and Educational Technology (LET) Research Lab, Faculty of Education and Psychology, P.O. Box 2000, University of Oulu, 90014 Oulu, Finland.
Email: joni.lamsa@oulu.fi

Funding information
Academy of Finland, Grant/Award Numbers: 292466, 311877, 318905, 331817

Abstract

Background: Cognitive load (CL) management is essential in safety-critical fields so that professionals can monitor and control their cognitive resources efficiently to perform and solve scenarios in a timely and safe manner, even in complex and unexpected circumstances. Thus, cognitive load theory (CLT) can be used to design virtual reality (VR) training programmes for professional learning in these fields.

Objectives: We studied CL management performance through behavioural indicators in authentic VR flight training and explored if and to what extent physiological data was associated with CL management performance.

Methods: The expert ($n = 8$) and novice pilots ($n = 6$) performed three approach and landing scenarios with increasing element interactivity. We used video recordings of the training to assess CL management performance based on the behavioural indicators. Then, we used the heart rate (HR) and heart rate variability (HRV) data to study the associations between the physiological data and CL management performance.

Results and Conclusions: The pilots performed effectively in CL management. The experience of the pilots did not remarkably explain the variation in CL management performance. The scenario with the highest element interactivity and an increase in the very low-frequency band of HRV were associated with decreased performance in CL management.

Takeaways: Our study sheds light on the association between physiological indicators and CL management performance, which has traditionally been assessed with behavioural indicators in professional learning in safety-critical fields. Thus, physiological measurements can be used to supplement the assessment of CL management performance, as relying solely on behavioural indicators can be time consuming.

KEYWORDS
cognitive load, cognitive load management, physiological measurements, professional learning, simulation, virtual reality

Virtual reality (VR) training systems are educational environments in which learners interact to mimic authentic situations that are relevant to the development of their professional expertise (Cook et al., 2013). Previous research has examined the kind of learning that occurs in VR training and has shown its effectiveness, for example, in the enhancing learners' technical performance in medicine (Chernikova et al., 2020). Although Radianti et al. (2020) noted that the use of VR systems was still in an experimental stage in many domains, the advantages of VR training systems have already become clear in safety-critical fields, where learners can face, familiarise themselves with, and solve scenarios that are rare, dangerous or difficult to create in real circumstances (Chernikova et al., 2020; Salas et al., 1998). In such complex learning, cognitive load theory (CLT) has been applied as a framework for designing training when the essential part of professional competence is cognitive load (CL) management (de Jong, 2010; Sweller, 1988; Sweller et al., 2019). When learners perform effectively in CL management, they can monitor and control their cognitive resources efficiently to perform and solve scenarios in a timely and safe manner, even in complex and unexpected circumstances (Aldekhyl et al., 2018). While CLT may help optimise VR training tasks design by redirecting the learners' limited working memory capacities from unnecessary CLs to cognitive processing of information that facilitates learning (Andersen & Makransky, 2021), the assessment of CL management performance may help to understand how learners succeed in redirecting these capacities themselves by engaging in CL management behaviours.

Elaborating on *how* (not only *what*) learning occurs during VR training from the perspective of CL management may assist instructors in providing proper guidance and developing further training for diverse needs, for example, professional development (experts) and acquiring the necessary professional expertise (novices). In addressing the question of *how* (professional) learning (Boshuizen et al., 2006) occurs, temporal analyses of learning (Lämsä et al., 2021), together with multimodal learning analytics (Olsen et al., 2020), have recently gained increasing attention. Here, mobile and wearable instruments, such as physiological measurement devices, have been used to provide indicators of the variation in the learners' CL with a high level of temporal granularity during VR training (Larmuseau et al., 2020; Minkley et al., 2021). Few studies, however, have focused on the associations between physiological indicators and CL management performance. The current study represents an attempt to better understand CL management in VR environments to support complex learning in safety-critical fields. We aim to examine the CL management of expert and novice pilots in the context of authentic VR flight training. To do this, we first use behavioural indicators to assess CL management performance, and then examine the extent to which physiological indicators are related to CL management performance.

1 | THEORETICAL FRAMEWORK

1.1 | Cognitive load theory

Since the 1980s, CLT has been widely used in attempts to design optimal learning and instruction situations (Sweller, 1988; see a recent example in Chen et al., 2021). CLT aims to 'explain how the information-processing load induced by learning tasks can affect students' ability to process new information and to construct knowledge in the long-term memory' (Sweller et al., 2019, pp. 261–262). Diverse sources can induce this information-processing load. The first source is intrinsic CL, which relates to element interactivity (Sweller et al., 2019). Element interactivity refers to the connections among the elements in a learning task that need to be perceived and performed (Sweller, 2010). Variation in element interactivity by task design can, therefore, also trigger variation in the learners' intrinsic CL. In the context of VR flight training, both the flying and monitoring pilots are required to collaborate to solve flight scenarios. This collaboration involves the sharing of subtasks and responsibilities, which helps to free up cognitive resources for both pilots to more efficiently solve the scenarios at hand (Janssen & Kirschner, 2020). As a result, effective collaboration can be used as an indicator of CL management performance in this type of professional learning context.

The second source is extraneous CL, which has traditionally been associated with the unfavourable design of instruction and activities that the instructional procedure requires learners to do, but that may not be necessary for learning (Kalyuga, 2011; Sweller et al., 2019). In immersive technology-enhanced settings, such as in VR training, researchers have seen extraneous CL as a multidimensional construct (e.g., Andersen & Makransky, 2021; Makransky et al., 2019): in addition to the instructions, the complex learning environment contributes to extraneous CL. For example, in authentic VR flight training systems that have a variety of instruments, the pilots must identify those instruments that provide relevant information in the given scenario (see Figure 1b in Section 2.1 as an example) to plan tasks effectively and manage time efficiently; that is, to perform well in CL management. Finally, although the interaction between the flying and monitoring pilots can decrease the intrinsic CL required for performing the task per pilot (see above), it also induces extraneous CL (Janssen & Kirschner, 2020) because of the coordination and communication processes between the pilots required for achieving and maintaining a shared understanding of the ongoing situation. The extent to which the interaction between the pilots decreases intrinsic CL and increases extraneous CL is associated with the pilots' CL management: for example, if the available cognitive resources are limited, the pilot can avoid the further increase of extraneous CL by ignoring the coordination and communication processes between the pilots and only delegating the tasks to the monitoring pilot.

In contrast to intrinsic and extraneous CL, germane CL does not induce overall CL, instead it relates to the internal cognitive processing of information by the learner (Sweller et al., 2019). For that reason, germane CL is not always distinguished as a separate source of CL, as are intrinsic and extraneous CL (Sweller, 2010), but is associated with intrinsic load (Kalyuga, 2011). Here, germane CL refers to the learner's ability to redistribute their cognitive resources from the extraneous to intrinsic aspects of the task, fostering their CL management and allowing them to use more resources for beneficial learning activities (Sweller et al., 2019).

Multiple studies have shown an association between CL and expertise (Kalyuga et al., 2003; 2010; Sweller, 1994). Because novices

must usually process several separate pieces of information (high element interactivity) and experts may only need to process a single or few elements (low element interactivity; Billett, 2001), the intrinsic CL of a learner with a high level of expertise (e.g. a pilot with long work experience) may be, on average, lower than that of a learner with less expertise (e.g. a pilot who has just acquired the necessary expertise for the pilot profession; Aldekhyl et al., 2018). Such differences between experts and novices can emerge as differences in their CL management performance, that is, how they monitor and control their cognitive resources to perform and solve scenarios in a timely and safe manner. In the following section, we elaborate on the behavioural indicators that are used to assess the CL management performance of pilots. After that, we justify how the physiological indicators could indicate CL management performance.

1.2 | Assessing CL management performance

Based on the recent pilot training framework (European Union Aviation Safety Agency, 2020, p. 77–81), CL management is one of the nine professional competencies that are to be targeted and assessed in the pilots' initial and continuous professional learning. The aim is for pilots to learn to control and monitor their cognitive resources to avoid cognitive overload that could hinder decision making and reduce safety margins during flight. The assessment of CL management performance consists of nine different overt behaviours (European Union Aviation Safety Agency, 2020, p. 81): (i) exercising self-control in all situations; (ii) planning, prioritising, and scheduling tasks effectively; (iii) managing time efficiently when carrying out tasks; (iv) offering and providing assistance; (v) delegating tasks when necessary; (vi) seeking and accepting assistance when necessary; (vii) reviewing, monitoring, and cross-checking actions; (viii) verifying that tasks are completed to the expected outcome; and (ix) managing and recovering from interruptions, distractions, variations, and failures effectively.

While these behavioural indicators provide a context-specific and detailed method to assess CL management performance, the monitoring of various indicators in real time or in a retrospective manner requires much time and resources and is vulnerable to subjective interpretations. To address the advantages of behavioural indicators, complementary—instead of compensatory—approaches are needed (Williges & Wierwille, 1979). With physiological measurement devices, researchers can capture components of the physical response of the human body to situations with varying CL (Naismith & Cavalcanti, 2015). For example, when a learner performs a complex task that represents a challenge to them, their bodily functions may change in response (Minkley et al., 2021). Changes in physiological parameters have been illustrated by previous research as being associated with varying CL (Ayres et al., 2021) in safety-critical fields, among others (Solhjoo et al., 2019).

Researchers have used various measurement devices to examine the associations between the variation in the CL and various physiological indicators (see Veltman & Gaillard, 1996), such as electrodermal activity (EDA; Johannessen et al., 2020), electroencephalography (EEG; Makransky et al., 2019; Parong & Mayer, 2021), heart rate (HR) and heart rate variability (HRV; Couceiro et al., 2019; Larmuseau et al., 2020) and pupil size (Szulewski et al., 2017). Usui and Nishida (2017) found that the decrease in very low-frequency (VLF) and high-frequency (HF) bands and increase in low-frequency (LF)/HF ratio of the HRV co-occurred during the increase in the intrinsic CL. Minkley et al. (2018) found that variations in HR and HRV were related to the complexity of the tasks that the participants were undertaking, with more complex tasks resulting in higher HR, a higher LF/HF ratio (see also Usui & Nishida, 2017) and lower root mean square of the successive differences between normal heartbeats (RMSSD); however, the differences in HR and RMSSD were not statistically significant. They also indicated that the physical response to task complexity might reveal a lack of resources regarding the demands of the task (Minkley et al., 2018), which indicates the potential for studying HR and HRV in expert–novice settings. Similarly, Kim and Jo (2019) found that prior knowledge was associated with variations in HRV during the learning process. Larmuseau et al. (2020) showed that HR variations seemed to indicate variations in the CL, and HR might indicate task difficulty more reliably than HRV, even though the authors could not explain much variance in the CL through the physiological data. Although the increase in HR and decrease in HRV have usually been assumed to indicate an increase in CL, there are individual differences in the humans' physiological responses (Tervonen et al., 2021). Solhjoo et al. (2019), for example, found that VLF and RMSSD increased with increasing intrinsic CL. An advantage of measuring HR/HRV is that participants do not typically feel that the HR/HRV measurements are intrusive and obtrusive (Mangaroska et al., 2021), and these measurements are less susceptible to motion, making them a flexible option for many settings (Zhou et al., 2020).

Altogether, HR and HRV, together with other listed physiological indicators, have shown preliminary potential for capturing variations in CL. In this study, we assume that variation in CL prompts adjustments in CL management behaviours. CL management performance is assessed using the behavioural indicators of the Evidence Based Training regulation (European Union Aviation Safety Agency, 2020, p. 81) in the context of authentic VR flight training. Thus, our aim is to study to what extent physiological indicators are associated with CL management performance. To address this aim, we propose the following research questions:

RQ1. How do the pilots perform in CL management in the VR flight training with increasing element interactivity?

RQ2. To what extent can CL management performance be explained in terms of physiological data in the VR flight training with increasing element interactivity?

2 | METHODS

2.1 | Participants and context

The participants were expert ($n = 8$) and novice pilots ($n = 6$), who were all male. The difference between the experts and novices was the amount of piloting experience. The experts had years of experience in piloting (at least 10 years). They were pilot trainers, and they practised in the simulator yearly. On the other hand, novices had just completed their simulator-based training and acquired the necessary professional expertise (Billett, 2001, section 1.1). On average, the participants performed a 47-min (SD = 9 minutes) VR flight training session in the Airbus A320 Full Flight simulator (Figure 1) that has been certified in the highest category (level D) of regulated flight training devices (European Union Aviation Safety Agency, 2012) providing full physical resemblance with a real aircraft including visual and motion systems. All the participants acted as flying pilots in the session, and as always, they had a monitoring pilot with them. All the participants were trained with the same experienced monitoring pilot. Participation was voluntary, and informed consent was obtained from all participants before participation.

The session had high authenticity in terms of both the VR training system and the structure of the training, during which the participants performed authentic scenarios that followed simple-to-complex whole-task sequencing (Van Merrienboer et al., 2003). In the present study, we focused on three approach and landing scenarios: on average, after 11, 31 and 42 minutes of flight (SD = 2, 9 and 9 min, respectively), the participants approached and landed with (i) a light wind, (ii) a heavy crosswind, and (iii) a heavy crosswind preceded by a fault situation. The professional trainers designed the scenarios so that in each scenario, the element interactivity was higher than in the previous scenario. We provide a detailed description of the VR flight training and how the element interactivity increased between the scenarios in supplementary online material. Between the scenarios, there was a 1- to 3-min pause, during which the simulator calculated the parameters and the pilots oriented for the next scenario. The training scenarios followed the typical structure of a VR flight training session for the pilots.

2.2 | Data

To address RQ1, we used the video recordings of the VR flight training sessions, and an experienced pilot trainer assessed the pilots' CL management performance in the three flight scenarios based on the overt behavioural indicators (European Union Aviation Safety Agency, 2020, p. 81, section 2.2). In this recent pilot training framework by the European Union Aviation Safety Agency (2020, pp. 77–81), CL management performance is one of the nine professional competencies that are assessed in the pilots' training. The CL management performance was determined based on how well and how often the relevant behavioural indicators were demonstrated by the participant when it was required. The relevant indicators were assessed on a 1—5 scale so that the higher scores demonstrated more effective and regular emergence of the behavioural indicators with safer outcome. A score of 5 refers to an outcome that significantly enhances safety; 4 refers to one that enhances safety; 3 refers to safe operation; 2 refers to an outcome that did not result in an unsafe situation; and 1 refers to an outcome that results in an unsafe situation. Not all the indicators were relevant in each scenario: when the

FIGURE 1 (a) Airbus A320 full flight simulator and (b) its cockpit.

participants approached and landed with (i) a light wind and (ii) a heavy crosswind, three indicators were observed (i.e., planning, prioritising and scheduling tasks effectively; reviewing, monitoring, and cross-checking actions; verifying that tasks are completed to the expected outcome; the subscale of these three indicators was reliable, Cronbach's $\alpha = 0.92$, 95% confidence interval = [0.85, 0.96]); in the case of (iii) a heavy crosswind preceded by a fault situation, seven indicators were observed (i.e., planning, prioritising and scheduling tasks effectively; managing time efficiently when carrying out tasks; delegating tasks when necessary; seeking and accepting assistance when necessary; reviewing, monitoring and cross-checking actions; verifying that tasks are completed to the expected outcome; and managing and recovering from interruptions, distractions, variations, and failures effectively; the subscale of these seven indicators was reliable, Cronbach's $\alpha = 0.85$, 95% confidence interval = [0.69, 0.94]).

To address RQ2, we averaged the assessments of the relevant behavioural indicators separately for each scenario and used these averaged assessments as the CL management performance in the subsequent analyses (14 pilots and each performed three scenarios, leading to the 42 assessments of CL management performance). Moreover, we collected the participants' physiological data using Firstbeat Bodyguard 2 measurement devices (Firstbeat, 2021), which have been suggested to achieve near medical-grade accuracy (Liu et al., 2022; Umair et al., 2021). We used the participants' artefact-corrected RR-interval data provided by the Firstbeat (Saalasti et al., 2005), from which we analysed their HR and HRV by relying on time- and frequency-domain parameters. For these calculations, we used the RHRV package in R (García Martínez et al., 2017).

2.3 | Analysis

To address RQ1, we visualised the participants' average CL management performance based on the relevant behavioural indicators for each scenario (see Section 2.2). We also conducted an analysis using generalised estimating equations (GEE, Liang & Zeger, 1986), which are extensions of generalised linear models for analysing clustered data (Liang & Zeger, 1986). In our study, we focused on three approach and landing scenarios, so we had three measurements from each participant (14 clusters, three measurements in each cluster). When using GEE, CL management performance in each scenario acted as the dependent variable while the scenario and the work experience of the pilot (expert or novice) acted as the independent variables. Since we had repeated measurements (three scenarios and measurements from each participant), we had to provide a correlation matrix to the model. We used the quasi-likelihood under the independence model criterion (QIC) for GEE (Pan, 2001) to select the most appropriate correlation matrix; we decided to use the auto-regressive correlation matrix because the QIC had the lowest value when applied, meaning that consecutive observations had a higher correlation than the correlation between the first and third observations. Next, we utilised bias-corrected GEE estimators for the regression parameters, which exhibit improved properties compared to standard GEE estimates, particularly when working with a small sample size ($N \approx 15$) (Lunardon & Scharfstein, 2017; Paul & Zhang, 2014). We conducted GEE and calculated the confidence intervals for the bias-corrected estimates using the BCgee R package (Lunardon & Scharfstein, 2017).

To study RQ2, we first conducted a baseline calculation based on our artefact-corrected RR-interval data. As a baseline, we used basic flying after the first take-off but before the beginning of the first scenario (Solovey et al., 2014). Second, we calculated the average of the following parameters over the baseline and three approach and landing scenarios:

1. HR
2. standard deviation of normal-to-normal intervals [SDNN, in milliseconds (ms)]
3. RMSSD (in ms), and

power of the following frequency bands of the HRV:

4. VLF in ms^2 (0.03–0.05 Hz),
5. LF in ms^2 (0.05–0.15 Hz),
6. HF in ms^2 (0.14–0.40 Hz), and
7. LF/HF.

We calculated the frequency-domain parameters using the wavelet transform (García Martínez et al., 2017). The duration of the baseline and each of the scenarios was 5 min. To facilitate the comparison among the pilots, we then normalised the averaged HR and HRV parameters using the following equation (Novak et al., 2015):

$$\text{indicator}_{\text{normalised}} = \frac{\text{indicator}_{\text{scenario}} - \text{indicator}_{\text{baseline}}}{\text{indicator}_{\text{baseline}}}.$$

Thus, the normalised value of the indicator shows how many percentages higher ($\text{indicator}_{\text{normalised}} > 0$) or lower ($\text{indicator}_{\text{normalised}} < 0$) the value of the indicator is compared with the baseline value. For further analyses, we transformed the HR and HRV parameters by taking the natural logarithm due to their skewed distributions. To examine the associations between CL management performance and physiological indicators, we used GEE. Again, CL management performance in each scenario acted as the dependent variable while the scenario, the work experience of the pilot (expert or novice), and HR and HRV parameters acted as the independent variables. We checked for multicollinearity among the HR and HRV parameters and found that a time-domain parameter (RMSSD) and two frequency-domain parameters (LF, HF) were highly correlated with other HR and HRV parameters. Therefore, these three parameters were excluded from further analysis. We used the auto-regressive correlation matrix and calculated confidence intervals for the bias-corrected estimates as in the analysis for RQ1. We also calculated the Wald test statistic to determine whether the full model with HR and HRV parameters improved the model (RQ2) compared to the model that only included the scenario and work experience of the pilot (RQ1).

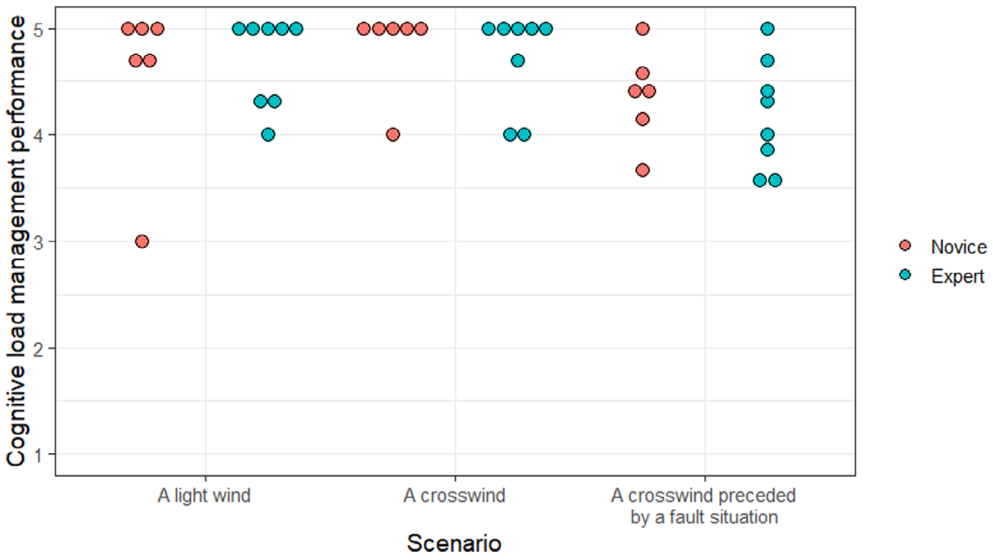

FIGURE 2 The CL management performance of the experts ($n = 8$) and novices ($n = 6$) in the three different approach and landing scenarios with increasing element interactivity.

TABLE 1 The results of the generalised estimation equations: (1) a reduced model in which the scenario and pilots' experience were used to explain the CL management performance (RQ1) and (2) a full model in which physiological data (heart rate [HR] and heart rate variability [HRV]) were added to explain the CL management performance (RQ2).

	Independent variable	Reduced model: CL management	Full model: CL management
Intercept		4.7 (4.2, 5.1)	4.5 (4.1, 4.9)
Scenario			
	A crosswind	0.12 (−0.04, 0.28)	0.06 (−0.08, 0.20)
	A crosswind preceded by a fault situation	**−0.39 (−0.64, −0.13)**	**−0.36 (−0.65, −0.07)**
Experience			
	Expert	−0.03 (−0.51, 0.45)	0.13 (−0.26, 0.52)
HR			
	HR		−2.29 (−4.83, 0.25)
HRV			
	SDNN		0.40 (−0.18, 0.98)
	VLF		**−0.43 (−0.79, −0.08)**
	LF/HF		0.20 (−0.01, 0.42)
Correlation parameter		0.65	0.60

Note: The correlation parameter is a measure of the correlation between the repeated measurements. We present the regression coefficients (β) along with their corresponding estimates and 95% confidence intervals.

3 | RESULTS

The findings for RQ1 (How do the pilots perform in CL management in the VR flight training with increasing element interactivity?) showed that the pilots performed effectively in CL management. Figure 2 provides a visualisation of the performance of the experts and novices in the three approach and landing scenarios with increasing element interactivity. When interpreting the results, it should be noted that scores 1 or 2 in the CL management performance were below or at the minimum acceptable level, respectively, which would have endangered the safe flying situation. Thus, these scores are rare in authentic pilot training. On average, the performance was 4.6 (SD = 0.5) in the first scenario with a light wind, 4.8 (SD = 0.4) in the second scenario with a crosswind, and 4.2 (SD = 0.5) in the third scenario with a crosswind preceded by a fault situation. The results of GEE shows that the work experience of the pilots did not significantly explain variation in CL management performance ($\beta = -0.03$, 95% confidence interval of $\beta = (-0.51, 0.45)$, see Table 1). The model indicates that the pilots performed slightly worse in CL management in the third scenario ($\beta = -0.39$,95% confidence interval of $\beta = (-0.64, -0.13)$, see Table 1) compared to the first scenario.

Figure 3 presents the variation in the selected HR and HRV parameters across different approach and landing scenarios. The results indicate that, on average, there were no significant differences observed among the three scenarios, nor between the expert and novice pilots regarding these parameters. In terms of RQ2 (To what

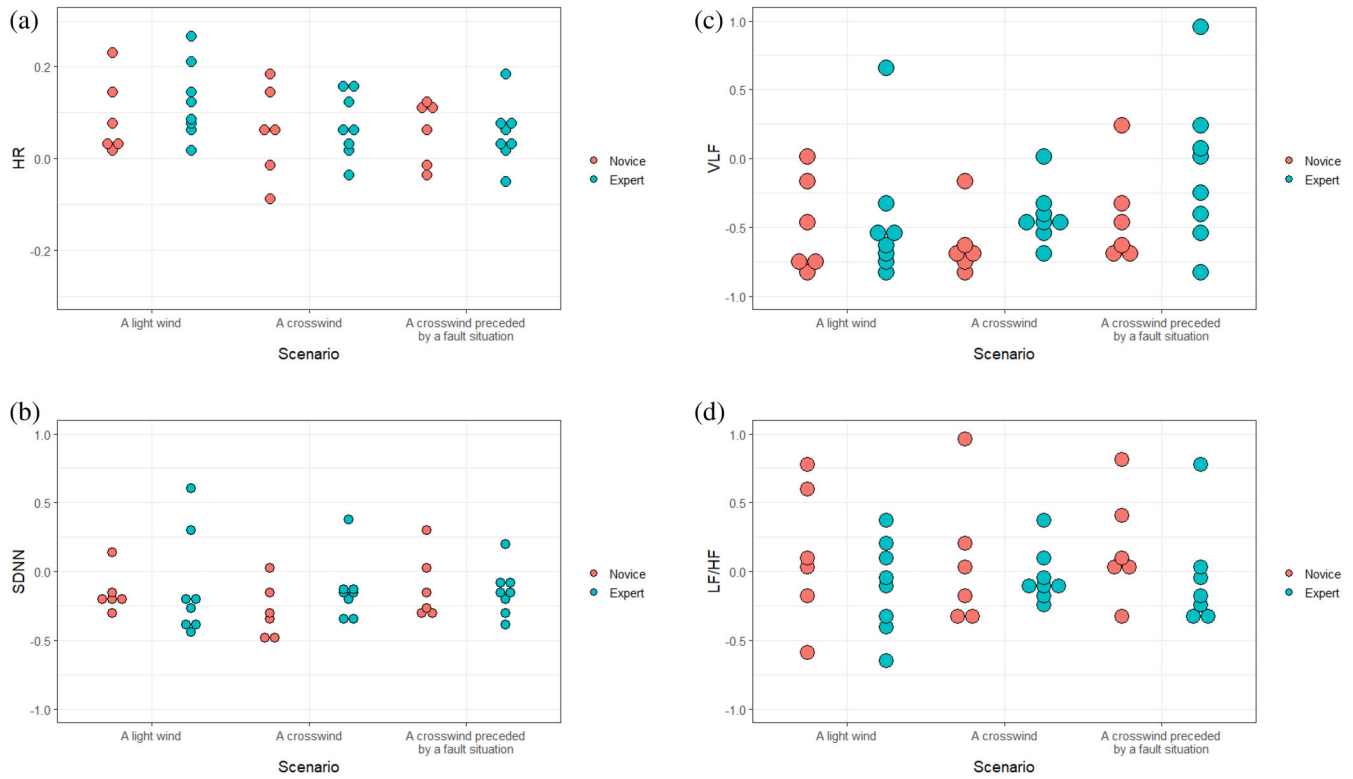

FIGURE 3 The normalised values of (a) heart rate (HR) and three HRV parameters for experts ($n = 8$) and novices ($n = 6$) in three different approaches and landing scenarios with increasing element interactivity. The HRV parameters include (b) the standard deviation of normal-to-normal intervals (SDNN), (c) very low-frequency band (VLF), and (d) low-frequency—high-frequency band ratio (LF/HF). The normalised values indicate the percentage increase or decrease in the parameter value compared to the baseline.

extent can CL management performance be explained in terms of physiological data in the VR flight training with increasing element interactivity?), we found that the HR and HRV parameters improved our model: We calculated Wald test statistic to compare the reduced model (RQ1) and full model (RQ2) and found that the full model was significantly better than the reduced model ($\chi(\text{df}=4) = 25.9, p < 0.001$). In the full model, the VLF band of the HRV was associated with CL management performance ($\beta = -0.43$, 95% confidence interval of $\beta = (-0.79, -0.08)$, see Table 1). When the VLF increased, CL management performance declined.

4 | DISCUSSION

In this article, first, we used behavioural indicators to assess CL management performance in the authentic VR flight training with increasing element interactivity. Second, we investigated to what extent HR and HRV parameters were associated with the CL management performance in the training. We found that the pilots performed effectively in CL management during three approach and landing scenarios, but the performance was slightly lower in the third scenario with the highest element interactivity (RQ1, Table 1). The lower performance in the scenario with the highest element interactivity related to lower scores in one or a few behavioural indicators of CL management performance (Section 2.2). For example, lower scores in the indicators related to the collaboration between the pilots occurred when pilots managed their CL by avoiding coordination and communication processes with the co-pilot that would have produced momentary extraneous CL (Janssen & Kirschner, 2020). Even though the collaboration between the pilots can distribute both pilots' cognitive resources to solve the scenarios at hand more efficiently, the pilots may have different dispositions towards the 'distribution advantage' (see Janssen & Kirschner, 2020, p. 787) and collaboration in general.

We found no significant differences between the experts and novices in CL management performance (RQ1, Table 1). Although the intrinsic CL was, on average, likely higher among the novices than the experts, the overall CL of the novices was not necessarily remarkably higher. Namely, the extraneous CL also depends on the situation in which the task is presented (Beckmann, 2010): Even though novices must usually process several separate pieces of information (high element interactivity and intrinsic CL), they might be even more familiar with the activities performed in the VR training than experts (cf. Paas et al., 2004), who suffered from a lack of set routines since they had had temporary layoffs because of COVID-19. Due to the complex VR flight training environment, both the experts and novices might also experience decreasing extraneous CL from the learning environment over time (e.g., it became easier to pay attention to the relevant parameters and instruments; see

Figure 1b), which illustrates the importance of following simple-to-complex task sequencing (Van Merrienboer et al., 2003) when the pilots can warm up and get used to the basic pilot procedures. In our study, we found that CL management performance remained stable or even slightly increased from the first to the second scenario (Table 1), even though element interactivity also increased from the first to the second scenario.

We also found that HR and HRV parameters improved the model that aimed to explain variation in the CL management performance (RQ2). Specifically, we found that better CL management performance was associated with a decrease in the VLF band of HRV (Table 1, see also Figure 3c). Decreasing VLF values (Solhjoo et al., 2019) have been linked to decreasing CL. However, it should be noted that the physiological basis of the VLF band is not fully understood (Kleiger et al., 2005; Shaffer & Ginsberg, 2017), and decreasing values of the VLF band have also been associated with increasing CL (Usui & Nishida, 2017). Therefore, it is possible that pilots who experienced more CL might place more emphasis on CL management behaviours and performed better in CL management. Usui and Nishida (2017) also found that the VLF band may recover slowly after an increase in intrinsic CL. In our study, the significance of the VLF band could also relate to the nature of the VR flight training: for example, the fault situation in the third approach and landing scenario only caused a momentary increase in the intrinsic CL that is challenging to capture with parameters that stabilise quickly after the fault has been solved.

When interpreting our findings, the following limitations should be considered: First, since we conducted our study in an authentic setting, we did not control all the variables. For example, VR flight training typically follows simple-to-complex task sequencing, so we did not counterbalance the order in which the scenarios were presented. Second, because of the unique context of our study (pilot training in Airbus A320 Full Flight simulator; see Figure 1), our sample size was small. In the future, with a larger sample size, the models explaining CL management performance could be developed and tested by splitting the data into training and test sets. Due to the small sample size and between-person variation in the physiological responses, the results and coefficients of the GEE in the full model (Table 1) should be interpreted cautiously (see Thomas et al., 2019). In the estimation, we took the small sample size into consideration by utilising the bias-corrected estimator that has proven to provide less biased estimates compared to the standard GEE estimates when the sample size is small (Paul & Zhang, 2014). Instead of providing evidence of the superiority of a specific HR or HRV parameter (such as VLF band), our results indicate that the physiological data, in general, has potential when assessing CL management performance during complex learning in VR environments. In the future, other physiological indicators (e.g. eye-tracking parameters; Johannessen et al., 2020) could be used to examine CL management performance in fields where an essential part of professional competence is to monitor and control one's own cognitive resources efficiently.

Despite these limitations, this study has several implications that can inform the development of VR training in safety-critical fields. First, our findings indicate that the VR training sessions (approximately 1–2 h) can be used for intensive practising because the CL management performance remained high throughout the session (RQ1). Second, it is important to practise basic procedures at the beginning of VR training and to avoid increasing unnecessary CL: namely because the CL management performance seemed to remain stable (or even improve slightly) from the first scenario to the second even though the element interactivity increased (Table 1). Third, besides monitoring CL with physiological measures in safety-critical fields, it is also essential to capture how learners monitor and control their cognitive resources efficiently to perform and solve scenarios in a timely and safe manner even in complex and unexpected circumstances, that is, how they perform in CL management. In doing this, context-specific and detailed assessment frameworks that utilise the expertise of trainers in the field are required. This study indicates that the HR and HRV measurements could be used to complement such assessment frameworks to inform learners and trainers on CL management performance.

Methodologically, we captured CL management performance with behavioural and physiological indicators in authentic VR training. One current challenge in CL research is the lack of reliable, unobtrusive measures for assessing CL in authentic and complex tasks, even though there are established sensitive measures of CL for simple and structured tasks (Ayres et al., 2021). This study contributes to this research gap, and despite its small sample size, provides preliminary evidence that the VLF band of HRV is related to CL management performance during complex learning. This finding is consistent with the previous research on the VLF band of the HRV (Solhjoo et al., 2019), even though Ayres et al. (2021) and Larmuseau et al. (2020) did not find HRV to be sensitive to variation in the CL. Our results can be useful in the development of instruments to capture CL management performance in professional learning. However, as the set of behavioural indicators (see Section 2.2) shows, CL management performance is a multidimensional construct. This multidimensional nature poses methodological challenges in capturing the temporal variation of CL management performance. For example, some behavioural indicators might be more sensitive to the variations in the different HR and HRV parameters, while variations in some of the HR or HRV parameters may not be sensitive enough to indicate CL management performance if caused by a sudden and quick event (e.g. the fault in the third scenario), although these types of events may cause a sudden increase in the need for information processing and even cognitive overload (Larmuseau et al., 2020), potentially leading to dramatic consequences in safety-critical fields.

5 | CONCLUSIONS

In the present study, we used a set of behavioural indicators to assess CL management performance in authentic VR flight training. Even though the work experience (expert–novice) of the pilots did not explain the variation in CL management performance, the scenario

with the highest element interactivity and increasing values of VLF band of the HRV related to lower performance. In the future, the models in which physiological data is used to predict CL management performance could be developed in two phases. First, a model with an adequate explanatory level should be developed with a training set. Second, the reliability and validity of the model should be studied with a test set. Capturing CL management performance is the first step to supporting learners so that they can monitor and control their cognitive resources efficiently to perform and solve scenarios in a timely and safe manner, even in complex and unexpected circumstances.

ACKNOWLEDGEMENTS

The work was supported by the Academy of Finland under Grant numbers 292466, 311877, 318905, and 331817. We greatly acknowledge Professor Dr. Matti Vihola and Senior Researcher Dr. Jouni Helske for their help in the data analysis. We also acknowledge Dr. Teuvo Antikainen for his initiative when creating this unique collaboration. We would also like to thank Tiina Kullberg and Max Lainema for their help in the data collection.

CONFLICT OF INTEREST STATEMENT

No potential conflict of interest was reported by the authors.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12817.

DATA AVAILABILITY STATEMENT

The data that support the findings of the current study are available on request from the corresponding author. The data are not publicly available because of privacy and ethical restrictions. The dataset is anonymous and stored in secure cloud services whose server rooms are located at the University of Jyväskylä, Finland. The metadata have been stored at the research portal of the University of Jyväskylä. When we conducted our study, we followed the guidelines of the Finnish Advisory Board on Research Integrity (2012). The Human Sciences Ethics Committee of the University of Jyväskylä made a positive statement regarding our study (755/13.00.04.00/2020) before we recruited participants and collected data. The authors have no potential conflicts of interest.

ORCID

Joni Lämsä https://orcid.org/0000-0001-7995-4090
Joonas Mannonen https://orcid.org/0000-0003-0285-9217
Ari Tuhkala https://orcid.org/0000-0001-5834-5595
Ville Heilala https://orcid.org/0000-0003-2068-2777
Ilkka Tynkkynen https://orcid.org/0000-0002-0347-3204
Emilia Lampi https://orcid.org/0000-0001-7660-3754
Katriina Sipiläinen https://orcid.org/0000-0002-3869-7320
Tommi Kärkkäinen https://orcid.org/0000-0003-0327-1167
Raija Hämäläinen https://orcid.org/0000-0002-3248-9619

REFERENCES

Aldekhyl, S., Cavalcanti, R. B., & Naismith, L. M. (2018). Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on Medical Education*, *7*, 23–32.

Andersen, M. S., & Makransky, G. (2021). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning*, *37*(1), 183–196. https://doi.org/10.1111/jcal.12478

Ayres, P., Lee, J. Y., Paas, F., & van Merrienboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, *12*, 702538. https://doi.org/10.3389/fpsyg.2021.702538

Beckmann, J. F. (2010). Taming a beast of burden – On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, *20*(3), 250–264. https://doi.org/10.1016/j.learninstruc.2009.02.024

Billett, S. (2001). Knowing in practice: Re-conceptualising vocational expertise. *Learning and Instruction*, *11*(6), 431–452. https://doi.org/10.1016/S0959-4752(00)00040-2

Boshuizen, H. P. A., Bromme, R., & Gruber, H. (Eds.). (2006). *Professional learning: Gaps and transitions on the way from novice to expert*. Springer. https://doi.org/10.1007/1-4020-2094-5

Chen, Y.-C., Chang, Y.-S., & Chuang, M.-J. (2021). Virtual reality application influences cognitive load-mediated creativity components and creative performance in engineering design. *Journal of Computer Assisted Learning*, *38*, 6–18. https://doi.org/10.1111/jcal.12588

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, *90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine*, *88*(6), 872–883 https://journals.lww.com/academicmedicine/Fulltext/2013/06000/Technology_Enhanced_Simulation_to_Assess_Health.34.aspx

Couceiro, R., Duarte, G., Durães, J., Castelhano, J., Duarte, C., Teixeira, C., Castelo Branco, M., Carvalho, P., & Madeira, H. (2019). Biofeedback augmented software engineering: Monitoring of programmers' mental effort. In *2019 IEEE/ACM 41st international conference on software engineering: New ideas and emerging results* (pp. 37–40). IEEE. https://doi.org/10.1109/ICSE-NIER.2019.00018

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

European Union Aviation Safety Agency. (2012). Certification specifications for aeroplane flight simulation training devices. https://www.easa.europa.eu/en/downloads/1735/en

European Union Aviation Safety Agency. (2020). Appendix to opinion No 08/2019 (A) (RMT.0599). https://www.easa.europa.eu/sites/default/files/dfu/Appendix%20to%20Opinion%20No%2008-2019%20%28A%29%20%28RMT.0599%29.pdf

Firstbeat. (2021). *Firstbeat Bodyguard 2: Accurate and reliable heart rate variability monitoring*. https://assets.firstbeat.com/firstbeat/uploads/2015/09/Firstbeat-Technologies-Bodyguard-2-Technical-Information.pdf

García Martínez, C. A., Otero Quintana, A., Vila, X. A., Lado Touriño, M. J., Rodríguez-Liñares, L., Rodríguez Presedo, J. M., & Méndez Penín, A. J. (2017b). *Heart Rate Variability Analysis with the R package RHRV*. Springer International Publishing. https://doi.org/10.1007/978-3-319-65355-6

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a

research agenda. *Education Technology Research and Development*, *68*(2), 783–805. https://doi.org/10.1007/s11423-019-09729-5

Johannessen, E., Szulewski, A., Radulovic, N., White, M., Braund, H., Howes, D., Rodenburg, D., & Davies, C. (2020). Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. *Computers in Human Behavior*, *111*, 106393. https://doi.org/10.1016/j.chb.2020.106393

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*(1), 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kim, J., & Jo, I.-H. (2019). Feasibility and use of psychophysiological responses based on cognitive load theory. *Australasian Journal of Educational Technology*, *35*(3), 150-165. https://doi.org/10.14742/ajet.4163

Kleiger, R. E., Stein, P. K., & Bigger, J. T. (2005). Heart rate variability: Measurement and clinical utility. *Annals of Noninvasive Electrocardiology: The Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc*, *10*(1), 88–101. https://doi.org/10.1111/j.1542-474X.2005.10101.x

Larmuseau, C., Cornelis, J., Lancieri, L., Desmet, P., & Depaepe, F. (2020). Multimodal learning analytics to investigate cognitive load during online problem solving. *British Journal of Educational Technology*, *51*(5), 1548–1562. https://doi.org/10.1111/bjet.12958

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.2307/2336267

Liu, S., Koch, K., Zhou, Z., Maritsch, M., He, X., Fleisch, E., & Wortmann, F. (2022). Toward nonintrusive camera-based heart rate variability estimation in the car under naturalistic condition. *IEEE Internet of Things Journal*, *9*(14), 11699–11711. https://doi.org/10.1109/JIOT.2021.3131742

Lunardon, N., & Scharfstein, D. (2017). BCgee: Bias-corrected estimates for generalized linear models for dependent data (0.1). https://CRAN.R-project.org/package=BCgee

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Lampi, E. (2021). What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review. *Educational Research Review*, *33*, 100387. https://doi.org/10.1016/j.edurev.2021.100387

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mangaroska, K., Martinez-Maldonado, R., Vesin, B., & Gašević, D. (2021). Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *Journal of Computer Assisted Learning, Advance Online Publication.*, *37*, 1030–1047. https://doi.org/10.1111/jcal.12542

Minkley, N., Kärner, T., Jojart, A., Nobbe, L., & Krell, M. (2018). Students' mental load, stress, and performance when working with symbolic or symbolic–textual molecular representations. *Journal of Research in Science Teaching*, *55*(8), 1162–1187. https://doi.org/10.1002/tea.21446

Minkley, N., Xu, K. M., & Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: Associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Frontiers in Education*, *6*, 632907. https://doi.org/10.3389/feduc.2021.632907

Naismith, L. M., & Cavalcanti, R. B. (2015). Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges*, *90*(11), 24–S35. https://doi.org/10.1097/ACM.0000000000000893

Novak, D., Beyeler, B., Omlin, X., & Riener, R. (2015). Workload estimation in physical human–robot interaction using physiological measurements. *Interacting with Computers*, *27*(6), 616–629. https://doi.org/10.1093/iwc/iwu021

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology*, *51*(5), 1527–1547. https://doi.org/10.1111/bjet.12982

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, *32*(1), 1–8. https://doi.org/10.1023/B:TRUC.0000021806.17516.d0

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125. https://doi.org/10.1111/j.0006-341X.2001.00120.x

Parong, J., & Mayer, R. E. (2021). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, *37*(1), 226–241. https://doi.org/10.1111/jcal.12482

Paul, S., & Zhang, X. (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in Medicine*, *33*(22), 3869–3881. https://doi.org/10.1002/sim.6198

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, *147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Saalasti, S., Seppänen, M., & Kuusela, A. (2005). Artefact correction for heart beat interval data. In T. Kärkkäinen & K. Majava (Eds.), *Advanced methods for processing bioelectrical signals: The first meeting of the Pro-Bisi project*. University of Jyväskylä https://assets.firstbeat.com/firstbeat/uploads/2015/11/saalasti_et_al_probisi_2004_congress.pdf

Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology*, *8*(3), 197–208. https://doi.org/10.1207/s15327108ijap0803_2

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00258

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Reports*, *9*, 14668. https://doi.org/10.1038/s41598-019-50280-3

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: Two field studies. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 4057–4066). Association for Computing Machinery. https://doi.org/10.1145/2556288.2557068

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

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

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L. A., & van Merrienboer, J. J. G. (2017). Measuring physician cognitive load:

Validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, *22*(4), 951–968. https://doi.org/10.1007/s10459-016-9725-2

Tervonen, J., Pettersson, K., & Mäntyjärvi, J. (2021). Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. *Electronics*, *10*(5). https://doi.org/10.3390/electronics10050613

Thomas, B. L., Claassen, N., Becker, P., & Viljoen, M. (2019). Validity of commonly used heart rate variability markers of autonomic nervous system function. *Neuropsychobiology*, *78*(1), 14–26. https://doi.org/10.1159/000495519

Umair, M., Chalabianloo, N., Sas, C., & Ersoy, C. (2021). HRV and stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback. *IEEE Access*, *9*, 14005–14024. https://doi.org/10.1109/ACCESS.2021.3052131

Usui, H., & Nishida, Y. (2017). The very low-frequency band of heart rate variability represents the slow recovery component after a mental stress task. *PLoS One*, *12*(8), e0182611. https://doi.org/10.1371/journal.pone.0182611

Van Merrienboer, J. J. G., 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://doi.org/10.1207/S15326985EP3801_2

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Williges, R. C., & Wierwille, W. W. (1979). Behavioral measures of aircrew mental workload. *Human Factors*, *21*(5), 549–574. https://doi.org/10.1177/001872087902100503

Zhou, T., Cha, J. S., Gonzalez, G., Wachs, J. P., Sundaram, C. P., & Yu, D. (2020). Multimodal physiological signals for workload prediction in robot-assisted surgery. *ACM Transactions on Human-Robot Interaction*, *9*(2), 1–26. https://doi.org/10.1145/3368589

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lämsä, J., Mannonen, J., Tuhkala, A., Heilala, V., Helovuo, A., Tynkkynen, I., Lampi, E., Sipiläinen, K., Kärkkäinen, T., & Hämäläinen, R. (2023). Capturing cognitive load management during authentic virtual reality flight training with behavioural and physiological indicators. *Journal of Computer Assisted Learning*, 1–11. https://doi.org/10.1111/jcal.12817