British Journal of Educational Technology *Vol 51 No 5 2020* 1548–1562
doi:10.1111/bjet.12958

Multimodal learning analytics to investigate cognitive load during online problem solving

Charlotte Larmuseau, Jan Cornelis, Luigi Lancieri, Piet Desmet and Fien Depaepe

Charlotte Larmuseau, PhD Candidate, Research member of ITEC, an Imec research group at KU Leuven; CRIStAL, research group at the University of Lille; and the Centre for Instructional Psychology and Technology (CIP&T). Technology-Enhanced Learning, Instructional Design, Physiological Data, (Multimodal) Learning Analytics. Jan Cornelis, Development Engineer and research assistant at Imec Leuven, Human Health, Wearable Human Health Technology, Physiology and Human Behaviour, CRISPR technology. Luigi Lancieri, Full Professor, head of NOCE team, Member of CRIStAL research center, University of Lille. Social Computing and Computer-Supported Cooperative Work (CSCW): Recommendation systems, Reuse of collective intelligence, Computer-Supported Creativity). Measure and analysis of Human factor in computer mediated Interactions (structure of social interactions, sentiments analysis, users' mobility behaviour, learning analytics). Piet Desmet, Full Professor. Director of ITEC, an Imec research group at KU Leuven. Academic director of the Imec smart education research program. Second Language Acquisition and Technology, with applications in Computer-assisted Language Learning (AI-based chatbots, prediction of linguistic complexity, intelligent feedback on complex learning tasks), Learning Analytics (effectiveness research, systems for adaptive & personalized testing and learning) and Language Technology and Corpus Linguistics (automated analysis and annotation of text corpora using natural language processing, parallel and multimedia corpora, learner corpora). Fien Depaepe, Associate professor at the Faculty of Pyschology and Educational Sciences of the KU Leuven, Research member of the Centre for Instructional Psychology and Technology (CIP&T) and principal investigator of ITEC, an Imec research group at KU Leuven. Instructional Design, Educational Effectiveness of Technoloy-Enhanced Learning. Address for correspondence: ITEC, Imec Research Group at KU Leuven campus Kulak, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. Email: charlotte.larmuseau@kuleuven.be

Abstract

To have insight into cognitive load (CL) during online complex problem solving, this study aimed at measuring CL through physiological data. This study experimentally manipulated intrinsic and extraneous load of exercises in the domain of statistics, resulting in four conditions: high complex with hints, low complex with hints, high complex without hints and low complex without hints. The study had a within-subject-design in which 67 students solved the exercises in a randomized order. Self-reported CL was combined with physiological data, namely, galvanic skin response (GSR), skin temperature (ST), heart rate (HR) and heart rate variability (HRV). Multiple imputation was used for handling missing data from resp. 16 and 19 students for GSR/ST and HR/HRV. First, differences between conditions in view of physiological data were examined. Second, we investigated how much variance of self-reported CL and task performance was explained by physiological data. Finally, we investigated which features can be used to assess (objective) CL. Results revealed no significant differences between the manipulated conditions in terms of physiological data. Nonetheless, HR and ST were significantly related to self-reported CL, whereas ST to task performance. Additionally, this study revealed the potential of ST and HR to assess high CL.

Practitioner Notes

What is already known about this topic

- Physiological data can be used to track changes in CL.
- GSR and HR(V) are most frequently used to measure CL unobtrusively.
- The effectiveness of these physiological measures for measuring CL remains inconclusive.

What this paper adds

- CL was systematically manipulated based on intrinsic and extraneous load.
- This paper explores the potential of ST for measuring CL.
- This paper combines physiological and self-reported data.

Implications for practice and/or policy

- ST and HR have the potential to measure high CL
- The perceived CL depends on internal and external conditions.
- Physiological data might not be sensitive to small differences in CL.

Introduction

In the search for a better understanding and supporting of learning, new manners of data collection such as sensing technology are explored to capture multimodal data unobtrusively in ecologically valid learning environments (Spikol & Cukurova, 2019). In the current study, we used physiological data to measure cognitive load during the online problem-solving process. Cognitive load was defined based on the Cognitive Load Theory (CLT) introduced by Sweller (1994). CLT indicates that cognitive load (CL) can be induced by both *intrinsic* and *extraneous* load (Sweller, 1994). The level of *intrinsic load* is determined by the amount of element interactivity and their interrelationships that need to be mastered by the learner. *Extraneous load* is mainly imposed by instructional procedures that induce unnecessary working memory load (Sweller, 2010). By monitoring task complexity and instructional support, personalized online courses can be developed to optimize CL. In view of optimizing CL, it is important to accurately measure CL during the online problem-solving process. Former studies used physiological measurements such as GSR, ST, HR and HRV to investigate CL (Cranford, Tiettmeyer, Chuprinko, Jordan, & Grove, 2014; Haapalainen, Kim, Forlizzi, & Dey, 2010; Larmuseau, Vanneste, Cornelis, Desmet, & Depaepe, 2019; Nourbakhsh, Wang, Chen, & Calvo, 2012). Despite the merits of these studies our current understanding of the association between physiological data and CL is characterized by at least three limitations. First, most studies did not systematically manipulate CL based on insights from CLT (Morton *et al.*, 2019). Second, the majority of studies did not combine physiological data with self-reports and task performance (Dindar, Malmberg, Järvelä, Haataja, & Kirschner, 2019; Larmuseau *et al.*, 2019). Third, there is no unambiguous answer to which physiological features are best in assessing CL (Larmuseau *et al.*, 2019).

To meet these limitations, the current study first experimentally manipulated the intrinsic load and the extraneous load by respectively varying the difficulty level of statistical exercises and the instructional support. More particularly, four sets of exercises were developed that differed on these two dimensions. In addition, students also received a computer-based operation span test (OSPAN) and participated in baseline measurements. OSPAN was used as verification of high CL, whereas the baseline measurement was a verification of low CL (Yuan, Steedle, Shavelson, Alonzo, & Oppezzo, 2006). We examined differences between the four sets, OSPAN and baseline

measurement, in view of the physiological data. Second, we combined physiological data with self-reported CL and task performance to investigate how much variance in these variables was explained by physiological data. Third, we investigated which physiological features were important for distinguishing high and low CL.

Theoretical framework

Cognitive load theory

CLT was introduced by Sweller (1994) to consider instructional implications on characteristics of human cognitive architecture, with a special focus on the limited capacity of the working memory. CL is the working memory load determined by the working memory resources required for performing a cognitive task (Kalyuga & Singh, 2016). A major goal of CLT is to optimally manage CL, since both overload and underload can lead to substandard performance (Chen *et al.*, 2016). CLT distinguishes between intrinsic and extraneous load. *Intrinsic load* is determined by the level of element interactivity. In case of higher element interactivity, the learning material is more complex requiring more intrinsic processing from learners' working memory for coordinating and integrating the learning material. *Extraneous load* refers to extraneous processing that does not contribute or even obstructs learning due to unnecessary mental demands (Sweller, van Merriënboer, & Paas, 2019). Instructional design techniques can reduce extraneous load by preventing learners to pay attention to irrelevant elements. For instance, instructional support can contribute to reducing this extraneous load by offering hints to tackle the learning tasks (Cierniak, Scheiter, & Gerjets, 2009; Sweller, 2010). In view of optimal learning and performance, an optimal level of CL is desirable (Sweller *et al.*, 2019). Therefore, it is important to accurately assess CL during the online problem-solving process to detect high complex learning material or suboptimal instructional procedures. Accordingly, personalized online learning environments can be developed that are adjusted to learners' working memory capacity level.

Measurement of cognitive load

A typical approach for measuring CL is through unidimensional rating scales (Chen *et al.*, 2016). In this respect, a frequently used rating scale is the Paas' (1992) nine-point mental effort rating scale (Chen *et al.*, 2016). This scale requires learners to rate their mental effort immediately after completing a task (Paas, 1992). This scale can be used as an index of overall CL (Chen *et al.*, 2016). Despite the merits of this rating scale, such as a quick administration of the perceived CL, some limitations should be mentioned. First, this scale requires learners to introspect on their cognitive processes which can induce biased results (Boekaerts, 2017). Second, these scales are obtrusive as they interrupt task flow. Third, those rating scales do not easily capture variations in load over time (Chen *et al.*, 2016). In contrast to self-reported data, physiological data can be measured at a high frequency and with a high precision during online problem solving (Di Mitri, Schneider, Specht, & Drachsler, 2018). Given the close relationship between CL and neural systems, the physiological approach is seen as a promising avenue to assess CL (Chen *et al.*, 2016). In the next section, we will explain how physiological data such as GSR, HR(V) and ST, can be unobtrusively measured by means of wrist-worn wearables and patches.

Galvanic skin response (GSR)

GSR, also known as skin conductance, refers to the variation of the electrical properties of the skin in response to sweat secretion (Benedek & Kaernbach, 2010). The time series of skin conductance can be characterized by a slowly varying tonic activity (ie, skin conductance level) and a fast varying phasic activity (ie, skin conductance responses; Braithwaite, Watson, Robert, & Mickey, 2015). The bulk of the skin conductance literature mainly reports associations with stress (Braithwaite *et al.*, 2015; Smets *et al.*, 2018). Nonetheless, more and more researchers also investigated the

relationship between GSR and CL (Chen *et al.*, 2016). Nourbakhs *et al.* (2012) captured GSR data from learners conducting arithmetic and reading tasks. The tasks differed in difficulty level, which relates to intrinsic load. Resp. four and three difficulty levels were distinguished for the arithmetic and reading tasks. Results of ANOVA of 13 and 16 participants (arithmetic and reading tasks) indicated that GSR significantly differed between task difficulty levels. By contrast, Shi, Ruiz, Taib, Choi, and Chen (2007) investigated 11 subjects when dealing with four tasks that differed in level of difficulty, but the results revealed insignificant differences across task difficulty levels for GSR. Similarly Larmuseau *et al.* (2019) investigated GSR for tasks that differed in terms of element interactivity. Results of 15 participants indicated no noticeable difference in GSR data between a high and low element interactivity task. Nonetheless, significant differences were found between GSR during the baseline measurement (ie, low cognitive processing) and during high complex problem solving. Overall, studies indicate that GSR increases when CL increases.

Heart rate (HR) and heart rate variability (HRV)

HR and HRV can be measured by a non-invasive electrocardiographic (ECG) method. HR averages the number of beats per minute, whereas HRV indicates small changes in the intervals between successive heartbeats. The majority of the studies have revealed that HR and HRV can be used to measure stress (Kim, Cheon, Bai, Lee, & Koo, 2018; Smets *et al.*, 2018). Other studies have also associated HR and HRV with cognitive demands, in which they have shown that an increase in HR and a decrease in HRV indicates higher CL. For instance, Taelman, Vandeput, Vlemincx, Spaepen, and Van Huffel (2011) collected ECG data of 43 undergraduates in a laboratory experiment during a high CL task (ie, doing complex arithmetic exercises) and a low CL task (ie, watching a relaxing movie). Results of pairwise comparison, reveal that HRV was significantly higher during the rest phase when compared with the high CL task. Additionally, Cranford *et al.* (2014) measured CL through HR in the context of chemistry. Findings of 12 participants suggested that problems that were intentionally designed to induce higher CL resulted in a larger increase of HR, compared to problems designed to induce lower CL. Finally, in the study of Brouwer, Hogervorst, Holewijn, and van Erp (2014), 35 participants solved different difficulty levels of the N-back task (ie, working memory capacity test) while recording HR and HRV. No significant effects were observed for HR and HRV, but trends in the data revealed that HR varied as a consequence of task difficulty. In summary, we can assume that HR and HRV are promising in detecting changes in CL.

Skin temperature (ST)

Previous studies also measured ST to indicate stress. Stress can induce peripheral vasoconstriction which causes a rapid, short-term drop in ST. Moreover, stress can also cause a more delayed skin warming, providing two opportunities to quantify stress (Herborn *et al.*, 2015; Karthikeyan, Murugappan, & Yaacob, 2012; Smets *et al.*, 2018). Little research has used ST to assess CL. As an exception, Haapalainen *et al.* (2010) collected data from multiple sensors, one tracking ST, in view of detecting CL. A total number of 20 subjects had to solve six tasks that differed in complexity. Using personalized machine learning techniques (ie, Naïve Bayes Classifier) to assess CL, they concluded that ST can be used to distinguish between low and high complex tasks. By contrast, the study of Larmuseau *et al.* (2019) observed no significant differences in ST between low and high complex tasks. In general, studies that used ST to investigate CL remain scarce. Consequently, it is unclear whether ST can be used to detect CL.

Shortcomings and research questions

Overall, there are some limitations in the current research field. First, not all studies systematically manipulated CL according to CLT (Morton *et al.*, 2019). Second, the majority of studies did not combine physiological data with self-reported CL (Dindar *et al.*, 2019; Larmuseau *et al.*,

2019). Moreover, it might be interesting to link physiological data with task performance across the different sets of exercises, as previous studies indicated that physiological adjustments ensure that task performance remains the same across different levels of complexity (Brouwer *et al.*, 2014; Iani, Gopher, & Lavie, 2004). Third, it remains unclear which features are most indicative for assessing CL. Against this background, following research questions are formulated:

RQ1: Do the different conditions (ie, sets of exercises, OSPAN and baseline measurement) result in differences in physiological data?

RQ2a: Can self-reported CL be explained in terms of physiological data across the different sets of exercises?

RQ2b: Can task performance be explained in terms of physiological data across the different sets of exercises?

RQ3: Which physiological features are important in assessing CL?

Method

Participants and study design

Participants were 67 adolescents (67.2% female and 32.8% male). The average age was 19.37. Of these 67 participants, 21 were in there last year secondary education and 46 were in the first two years of higher education. Participation was voluntarily and all participants signed an informed consent before participation. A requirement for participation was that participants had been recently introduced to the theory on probabilistic reasoning in statistics. A within-subject design was used in which participants solved four different sets of exercises in a randomized order. The intervention is illustrated in Figure 1. Participants started with OSPAN. Afterward, baseline measurements were conducted where students watched a relaxing movie with headphones. Additionally, participants randomly received (to counteract a sequence effect) four sets of exercises (section 3.4.). Each session ended automatically. Within the Moodle Learning Management System (LMS) all answers were completed online and students were allowed to use a calculator on their computer. No information about the correctness of the answers was provided. After each set of exercises participants had to indicate their perceived CL (Paas, 1992) (Q*).

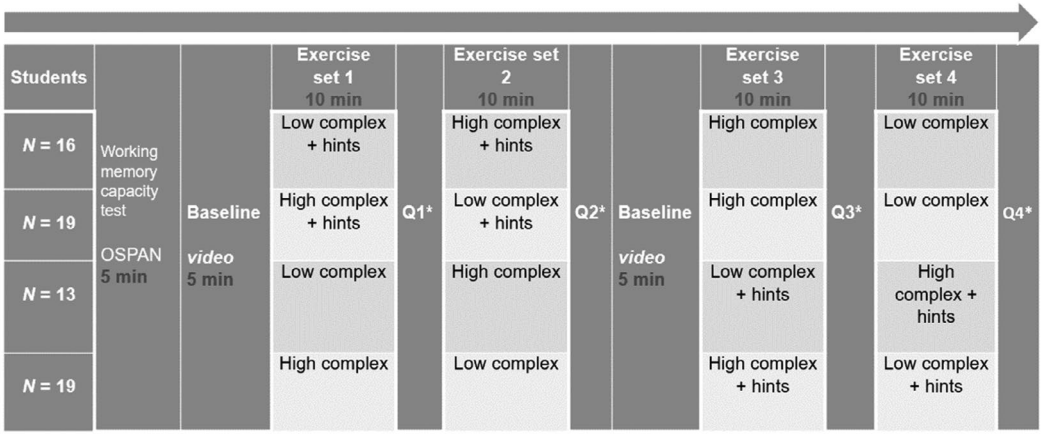

Figure 1: The study design

Physiological recordings

During the sets of exercises physiological data were measured by two wearables developed by Imec as indicated in Figure 2. The first wearable was a chest patch recording ECG at a sampling rate of 256 Hz and providing information about HR(V). The second wearable was a wrist-worn device (worn on the non-writing hand) measuring GSR with a high dynamic range (0.05–20 μS) at the lower side of the wrist. The output was accurate within a frame of approximately 1 second. ST was acquired at the upper side of the wrist at a frequency of 1 Hz. The output was accurate within a frame of approximately 1 second at 0.1°C. Both wearables measured the magnitude of acceleration. Using high quality physiological signals, 19 features were calculated (eight GSR features, four ST features and seven ECG features) in a window of 1 minutes. The quality of the data was automatically checked by an algorithm that calculated a quality indicator and took anomalies into account. Eventually there was a high drop-out of some wearables due to unsuccessful and low quality measurements: bugs in wearables, bad connection due to incorrect fitting, too dry skin, chest hair (ECG patches), etc. Only complete datasets (across all conditions) were retained, resulting in data from 51 wrist-worn wearables (GSR and ST) and 48 ECG patches (HR and HRV).

OSPAN

A computer-based operation span test (OSPAN; Stone & Towse, 2015) was used in which each participant was shown a number for 1s on a screen that he/she had to remember in the correct order. After each number, participants were shown a mathematical operation and had to decide on the correctness of the given answer. Accordingly, for every storage element (number to remember) there was a processing phase (a mathematical operation) immediately succeeding it. In total there were three experimental trails of each span size (number of digits to be remembered: 2–6). Since this dual-task requires both storage and the processing of the information, it is effortful and demanding of the working memory (Yuan *et al.*, 2006). Consequently, this test can be used as a control intervention for high CL.

Exercises and task performance

The content of the exercises was probabilistic reasoning in statistics. The four sets of exercises that were developed differed from each other on two dimensions. The first dimension related to intrinsic load. The complexity of the exercises was manipulated on the basis of element interactivity (Sweller, 2010). As illustrated in Figure 3, solving the low complex set of exercises consisted of applying one procedure or rule while the high complex exercise series consisted of several solution steps consisting of different elements (elements are in bold). The second dimension related to extraneous load and was manipulated by providing step-by-step instructions (ie, hints) that were offered just-in-time and could be consulted voluntarily (in italics in Figure 3). Consequently, we had four sets of exercises: (1) high complex with hints, (2) high complex without hints, (3) low

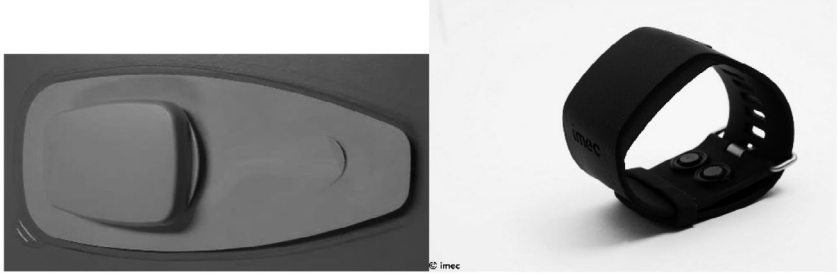

Figure 2: The Imec chest patch and Chillband

Low element interactivity	High element interactivity
Exercise: In a class, 5 children have to read. In how many different ways can you create a schedule with the order in which they have to read aloud? *Procedural information* ***(one element):*** *- Apply **the permutation rule** (order is important)*	**Exercise:** You have 8 men and 10 women. You want to form a jury of 12 members, but you want to have more men than women in the jury. How many unique combinations are there? *Procedural information* ***(three elements):*** *- Apply **the combination** rule (no strict order and no repetition)* *- Calculate situation 1 (7 men/5 women) and apply **the product rule*** *- Calculate situation 2 (8 men/4 women) and apply the product rule* *- **Add up** the total possibilities*

Figure 3: Example of an exercise of low and high element interactivity with hints

complex with hints and (4) low complex without hints. By systematically manipulating CL, we aimed at detecting differences in CL through physiological data. Task performance was retrieved from the total number of exercises that was correctly solved per session. All exercises had a precise correct answer and were scored as correct (1) or incorrect (0).

Analysis

Since features of physiological data might be strongly related, principal component analysis (PCA) was used to reconstruct the original dataset of 19 features. By conducting PCA, 7 features were extracted based on their standardized loadings after varimax rotation, namely: GSR_SCmag, ECG_sdnn, ST_mean, ECG_LF, LF_VLF_Ratio, ST_slope and ECG_meanHR. Detailed information about these features such as meaning and calculation, can be found in Table S1, in which the selected features are in bold. The purpose of the standardization was to facilitate individual-difference comparisons and thus, to factor out issues (eg, the thickness of the skin; Braithwaite *et al.*, 2015).

In view of *RQ1,* we investigated the mean differences in physiological data between the four sets of exercises, OSPAN and baseline measurement. Since the first baseline measurement was strongly influenced by the measurement of physiological data during the OSPAN (indicated by visual analysis), we opted to only use the second baseline measurement for analysis. We controlled for the magnitude of acceleration, as movement can influence physiological signals (Boucsein, 2012). A within-subjects effects of the different interventions by 6 × 1 repeated measure ANOVA was conducted. As multiple comparisons were conducted, Bonferroni correction was applied (Bretz, Hothorn, & Westfall, 2010).

Regarding *RQ2,* we investigated whether self-reported CL and task performance can be explained in terms of physiological data across the different sets of exercises. Multilevel Modeling (MLM) was applied to investigate how much variance of the self-reported data and task performance was explained by the physiological data. A first null model included the intrinsic manipulation and the extraneous manipulation. The subsequent analyses had perceived CL and task performance as dependent variables and the selected physiological features (ie, based on PCA) as predictors. The average variance explained by the predictors (r^2) is presented for each model. We also checked if the model was statistically better after inclusion of physiological data. To maintain statistical power while comparing the models, we decided to complete missing data via multiple

Table 1: Overview of the within-subjects effects of the different interventions

	GSR		*HR*		*HRV*						*ST*			
	GSR_SCmag		*ECG_meanHR*		*ECG_SDNN*		*ECG_LF*		*LF_VLF_Ratio*		*Mean ST*		*ST slope*	
I–J	*MD*	*SE*	*MD*	*SE*	*MD*	*SE*	*MD*	*SE*	*MD*	*SE*	*MD*	*SE*	*MD*	*SE*
Base–HC	−0.03	0.17	**−0.25****	0.07	0.04	0.15	0.02	0.06	−0.30	0.17	−0.02	0.06	−0.37	0.16
Base–HC + hints	−0.14	0.17	**−0.36****	0.07	0.22	0.15	−0.002	0.06	−0.16	0.17	0.04	0.06	−0.43	0.16
Base–LC	−0.06	0.17	−0.18	0.07	0.23	0.15	0.06	0.06	−0.00	17	−0.03	0.06	−0.10	0.16
Base–LC + hints	0.12	0.12	**−0.36****	0.07	0.23	0.15	−0.02	0.06	−0.36	0.17	0.04	0.06	−0.21	0.16
Base–OSPAN	−0.19	0.17	**−0.88****	0.07	**0.46***	0.15	0.04	0.06	−0.14	0.16	**0.45****	0.06	**−1.2****	0.16
HC–HC + hints	−0.14	0.17	−13	0.07	0.18	0.15	−0.04	0.06	0.14	0.17	0.06	0.06	−0.06	0.17
HC–LC	−0.08	0.17	0.06	0.07	0.03	0.15	0.04	0.06	0.29	0.17	−0.02	0.06	0.27	0.16
HC–LC + hints	0.09	0.17	−0.11	0.07	0.19	0.15	−0.04	0.06	−0.06	0.17	0.05	0.06	0.16	0.16
HC–OSPAN	−0.21	0.17	**−0.64****	0.07	0.42	0.15	0.02	0.06	0.16	0.17	**0.47****	0.06	**−1.4****	0.17
HC + hints–LC	0.06	0.17	0.06	0.07	−0.15	0.15	0.08	0.06	0.15	0.17	−0.07	0.06	0.33	0.16
HC + hints–LC + hints	0.23	0.17	0.02	0.07	0.01	0.15	0.00	0.06	−0.20	0.17	−0.00	0.06	0.22	0.16
HC + hints–OSPAN	−0.08	0.17	−0.51**	0.07	0.24	0.14	0.06	0.06	0.02	0.17	**0.41****	0.06	**−1.4****	0.16
LC–LC + hints	0.17	0.17	0.17	0.07	0.16	0.15	−0.08	0.06	−0.36	0.17	0.07	0.06	−0.11	0.16
LC–OSPAN	−14	0.17	**−0.70****	0.07	0.39	0.14	−0.02	0.06	−0.14	0.16	**0.48****	0.06	−1.7	0.16
LC + hints–OSPAN	−0.12	0.17	**−0.53****	0.07	0.23	0.14	0.06	0.06	0.22	0.16	**0.42****	0.06	**−1.6****	0.16

MD (mean difference = I–J)–Base, baseline measurement; HC, high load; LC, low load.
**< 0.01; *< 0.05.

imputation which resulted in data of 64 students. Creating multiple imputations, as opposed to single imputations accounts for the statistical uncertainty in the imputations and yields accurate standard errors (Azur, Stuart, Frangakis, & Leaf, 2011). Accordingly, this method can be used for a large amount of missing values such as ±30% missing observations in the current study (Fichman & Cummings, 2003).

With respect to *RQ3,* we investigated whether CL can be detected through physiological data. Informed by the results of RQ1, we only maintained the two interventions that significantly differed and conducted a binary classification, that is, baseline measurement (low CL) and OSPAN (high CL). The selected features, based on PCA to avoid overfitting, were incorporated in the machine learning model. Using 128 observations from 64 students we trained a logistic regression model in a 10-fold cross-validation approach (Ramasubramanian & Singh, 2016).

RESULTS

RQ1 investigated the differences between the four sets of exercises, OSPAN and the baseline measurement in terms of physiological data as indicated in Table 1. First, regarding *GSR* ($N = 51$), no significant differences between the six interventions were observed. Second, mean *HR* was significantly lower during the baseline measurement compared to the high complex set, high complex set with hints, low complex set with hints and OSPAN. Furthermore, *mean HR* ($N = 48$) was significantly lower during the four sets of exercises when compared with OSPAN. In terms of *HRV* ($N = 48$), results of SDNN (ie, standard deviation of the interbeat intervals) reveal that HRV was significantly higher during OSPAN when compared with the baseline measurement. Third, results of *ST* ($N = 51$) reveal that mean ST during OSPAN was significantly lower compared to all other interventions. Similarly, the slope of ST was higher during OSPAN when compared with the remaining interventions.

The results for *RQ2* are presented by Table 2. *RQ2a* investigated whether *self-reported CL* across the different sets of exercises can be explained through physiological data. First, in the baseline model, it was observed that the sets with high element interactivity resulted in higher self- reported CL, whereas the provision of hints reduced self- reported CL. Furthermore, in terms of physiological data, results indicated that CL is explained by a significant negative effect of the slope of ST and a positive effect of mean HR. The total variance explained is $r^2 = .18$ and the model including physiological data was significantly better: $\chi^2(7, N = 64) = 32.19$, $p < .001$ when compared with the baseline model. *RQ2b* investigated whether *task performance* across the four sets of exercises can be explained through physiological data. The baseline model revealed that high task complexity decreased task performance, whereas no effect of the provision of hints was found. In terms of physiological data, results reveal that a negative slope of ST is associated with higher task performance. The total variance explained is $r^2 = 0.47$ and the model including the physiological data was significantly better: $\chi^2\ (7, N = 64) = 15.31$, $p < .05$.

RQ3 aimed at identifying features that assess high (ie, OSPAN) and low CL (ie, baseline measurement) by constructing a logistic regression model. Table 3 reveals the importance of the slope of ST and mean HR for distinguishing between low and high CL. Higher values of mean HR and the slope of ST indicate higher (objective) CL. The overall accuracy, sensitivity and specificity were respectively 0.76, 0.74 and 0.79 indicating a good performance of the logistic regression model.

Discussion

Main findings and implications

This study aimed at measuring CL through physiological data. Therefore, this study systematically manipulated intrinsic and extraneous load to investigate how physiological features, namely,

Table 2: Multilevel analyses with self-reported CL and task performance as dependent variables (N = 64)

	Self- reported CL		*Task performance*	
	Est. (SE)	*Est. (SE)*	*Est. (SE)*	*Est. (SE)*
Intercept	−0.19(0.11)	−0.24(0.11)	0.62(0.09)	0.60(0.08)
Manipulations				
Complexity (High)	0.57(0.09)***	0.58(0.09)***	−1.34(0.08)***	−1.32(0.08)***
Hints	**−0.21(0.09)***	−0.20(0.09)*	0.07(0.08)	0.07(0.08)
GSR				
GSR_SCmag		0.01(0.09)		0.04(0.07)
HR and HRV				
ECG_meanHR		**0.26(0.07)*****		−0.03(0.06)
ECG_SDNN		0.00(0.06)		0.10(0.05)
ECG_LF		0.03(0.09)		−0.05(0.07)
LF_VLF_Ratio		−0.11(0.08)		0.07(0.05)
ST				
ST_mean		0.13(07)		0.03(0.06)
ST_slope		**−0.14(0.06)***		−0.15(0.05)**
Random effects	Variance (SD)	Variance (SD)	Variance (SD)	Variance (SD)
Participant	0.46(0.68)	0.34(0.59)	0.18(0.43)	0.17(0.42)
Residual	0.50(0.71)	0.50(0.70)	0.37(0.61)	0.37(0.61)
r^2	0.09	0.18	0.44	0.47

***< 0.001; **< 0.01; *< 0.05.

Table 3: The coefficients table of the logistic regression model (high CL; N = 124)

	Estimate	*SE*	*p*
(Intercept)	0.05	0.40	.87
GSR			
GSR_SCmag	0.01	0.19	.53
HR and HRV			
ECG_meanHR	**0.47**	0.13	.00***
ECG_SDNN	−0.14	0.12	.24
ECG_LF	0.12	0.19	.53
LF_VLF_Ratio	−0.15	0.11	.20
ST			
ST_mean	0.02	0.12	.88
ST_slope	**0.61**	0.12	.00***

***< 0.001; **< 0.01; *< 0.05.

GSR, ST and HR(V) vary as a result of changes in CL. More particularly, four sets of exercises on probability calculations in statistics were developed (1) high complex set with hints, (2) low complex with hints, (3) high complex without hints and (4) low complex without hints. Additionally, a working memory capacity test (ie, OSPAN) and baseline measurement (ie, watching a relaxing movie) was used as verification of resp. high and low CL.

RQ1 investigated whether the different conditions (ie, sets of exercises, OSPAN and baseline measurement) result in differences in physiological data. Trends in our data revealed that *GSR* increased with increased task difficulty, but no significant differences were observed. Our findings are in line with Larmuseau *et al.*'s study (2019). Nonetheless, our findings are in contrast with

the study of Nourbakhs *et al.* (2012) indicating that GSR reveal differences in task complexity. These contradictory findings can be explained by the difference in task design between the study of Nourbakhs *et al.* (2012) and our study. More particularly in Nourbakhs *et al.*'s (2012) study students were given a working memory task such as OSPAN in our study, in which the difficulty level increased (eg, the addition of more complex numbers). From the experience of our study, we can assume that cognitive activity does indeed increase as more information needs to be remembered and processed. The advantage of this type of tasks is that the students' prior knowledge is less important, which means that the CL for the students increases more evenly compared to the statistical exercises in our study. With respect to *HR*, our findings revealed that mean HR was significantly lower during the baseline measurement compared with the other interventions, with the exception of the low complex set of exercises without hints. These results might indicate that providing hints during the low complex set of exercises has provoked additional CL since instructional support might not have been necessary and might even have distracted students (Sweller, 2010). Similarly, results indicated that HR was significantly higher during OSPAN compared to the sets of exercises and the baseline measurement. This might suggest that OSPAN induced more CL than the other interventions. Similarly, Cranford *et al.* (2014) revealed that HR is higher in case of higher complexity levels. In our study, *HRV* (ie, SDNN) distinguished the OSPAN and baseline measurement. The fact that a decrease of HRV is associated with high CL was also observed by Taelman *et al.* (2011) who also used HRV to distinguish between a high complex task (ie, also a working memory task) and a rest phase (ie, also a relaxing movie). However, it seems that HR is a more reliable indicator of task difficulty when compared to HRV, which was also observed in the study of Brouwer *et al.* (2014). Regarding *ST*, our findings indicated that mean ST was lower and the slope of ST was higher during OSPAN when compared with the other interventions. However, ST did not vary as a result of the manipulated exercises. This latter finding is in line with the study by Larmuseau *et al.* (2019) where no differences were detected between a high and low complex task manipulated based on element interactivity. Summarized, against our expectations, the manipulation of the level of complexity of a task, based on two dimensions, that is, intrinsic (ie, element interactivity) and extraneous load (ie, provision of hints) did not result in differences in physiological data. As aforementioned, possible explanation for this is that we did not take important student characteristics into account (eg, prior knowledge). On the contrary, it is also possible that this is due to both the task and study design. For instance, students were allowed to choose whether or not to consult the hints. In addition, the study was also non-committal, which meant that students who found the exercises too difficult did not try to solve them. Accordingly, those students did not experience CL, which biased the results. Although our findings were not as expected in terms of the developed set of exercises, we observed that OSPAN induced high CL. In this study, OSPAN was used as a control intervention for continuous (high) CL as this dual-task requires both storage and processing of information (Yuan *et al.*, 2006). Moreover, as all physiological features are also related to stress, OSPAN might also have induced stress as a reaction on continuous high levels of CL (Herborn *et al.*, 2015; Iani *et al.*, 2004). OSPAN consisted of different trails of remembering and processing information. The more numbers students had to remember and the more mathematical problems they had to solve, the harder it was for students to complete the test correctly which might have induced both high CL and stress. This would be in accordance with the literature since task failure and feelings of lack of control have been shown to induce stress (Conway, Dick, Li, Wang, & Chen, 2013).

RQ2 examined to what extent variance in self-reported CL and task performance can be explained through physiological data. In terms of *self-reported CL* (RQ2a), results revealed a positive influence of the level of complexity and a negative influence of the provision of hints which is in line with CLT and indicates that we have succeeded in our manipulations (Sweller, 2010). Results

furthermore indicated that a less negative slope of ST is related to the invested CL. Based on the results of RQ1 (ie, positive slope during OSPAN) this might indicate that high levels of CL might have induced stress. Therefore, we can assume that students who experienced less stress were able to invest more CL. This hypothesis is in line with CLT stating that exceeding the working memory capacity (ie, cognitive overload) can interfere with learning (van Merriënboer & Sluijsmans, 2009; Sweller, 2010). Results also revealed that self-reported CL is positively related to higher HR, which is in line with the findings of Cranford *et al.* (2014). However, our findings also revealed that not much variance in CL was explained by physiological data. This might suggest that self-reported and physiological data provide different information about the learning process. In fact, physiological data relate to reactions that happen largely unconsciously whereas self-reported data are more conscious (Dindar *et al*, 2019). Nonetheless, the combination of both types of data provides more insight into the meaning of physiological data. In terms of *task performance* (RQ2b), our findings indicated that high levels of complexity negatively influenced task performance. No influence of the provision of support was observed. Again, not much variance in task performance was explained by physiological data. Results showed a negative influence of the slope of ST on task performance. When comparing results of task performance with the results of self-reported CL, we can infer that perhaps less stress led to better task performance. What is striking here is that HR and HRV had no influence on task performance. Presumably, students who were more relaxed during the problem-solving process (ie, less CL and stress) might also had better knowledge of the content and therefore, attained higher task performance.

RQ3 aimed at identifying the most important features for assessing high and low CL. Based on the findings of RQ1, we decided to compare OSPAN (ie, high CL) with the baseline measurement (ie, low CL). Results indicated that higher values in the slope of ST and mean HR are indicative for high CL. These results are in line with findings of RQ1 and RQ2, and further provide evidence that OSPAN required high levels of CL. When we take into account that OSPAN might also have induced stress, results are in line with the study of Karthikeyan *et al.* (2012). This study also used a working memory test (ie, Stroop color word test) to investigate whether ST can detect stress. Results revealed that ST was a reliable measure for identifying stress level changes. These findings demonstrate the potential of physiological signals to detect high CL or CL overload (ie, exceeding working memory capacity). Consequently, we have to take into account that in the current study stress might have obscured the relationship between CL and physiological data.

Limitations and suggestions for future research

Regardless of the fact that this study provides insight into the use of physiological data for measuring CL, the study is also characterized by some limitations. A first limitation relates to the fact that there was a lot of missing data which possible affected our results. A second limitation relates to the fact that we did not assess self-reported CL during OSPAN and the baseline measurement. This could have provided a clear insight into learners' perception of CL during these phases and further unravel the association between self-reported and physiological data. A third limitation was the noncommittal nature of the study, and consequently, this might has resulted in students being unmotivated to solve the exercises. Therefore, it might be useful for future research to integrate the study into the regular curriculum. Fourth, important internal conditions should not be overlooked in order to obtain meaningful information from multimodal data (Gašević, Dawson, & Siemens, 2015). Students' prior knowledge might have impacted our findings, as domain knowledge (eg, formula) has a major influence on CL (Sweller *et al.*, 2019). Also affective characteristics, such as negative feelings towards the learning material can induce CL by disrupting working memory processes, particularly for high complex tasks (Basanovic *et al.*, 2018). Therefore, this might have influenced the physiological recordings. A final limitation

is related to the machine learning model. The sample size was rather low for using machine learning techniques. Consequently, based on our data we cannot make major statements about the results. Therefore, it should be emphasized that it is mainly an indication of feature importance.

Conclusion

Against our expectations, results revealed that physiological data could not be used to detect differences in CL based on intrinsic and extraneous manipulations. By contrast, most of the significant results are related to OSPAN and the baseline measurement. Based on our findings related to OSPAN, we might be able to conclude that HR, HRV and ST is more sensitive to high CL, namely, exceeding the learner's cognitive capacity and the related mental states (ie, stress). In this respect, as high CL can also provoke stress, it is not always clear what exactly is measured via physiological data. Therefore, it remains important to combine self-reports of associated mental states with physiological data in future studies as this might facilitate interpretation.

Acknowledgements

This study was carried out within imec's Smart Education research program, with support from the Flemish government.

Statements on open data, ethics and conflict of interest

The dataset is anonymous and stored on the local drive of the computer of the main researcher and can be requested after approval of all authors.

This study design was approved by the ethical commission (G-2019 03 1605).

The authors declare no competing interests.

References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Basanovic, J., Notebaert, L., Clarke, P. J. F., MacLeod, C., Jawinski, P., & Chen, N. T. M. (2018). Inhibitory attentional control in anxiety: Manipulating cognitive load in an antisaccade task. *PLoS ONE*, *13*. https://doi.org/10.1371/journal.pone.0205720

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *1*, 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Boekaerts, M. (2017). Cognitive load and self-regulation: Attempts to build a bridge. *Learning and Instruction*, *51*, 90–7. https://doi.org/10.1016/j.learninstruc.2017.07.001

Boucsein, W. (2012). *Electrodermal activity*, 2nd ed. New York: Springer.

Braithwaite, J., Watson, D., Robert, J., & Mickey, R. (2015). *A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments (Revised version 2.0)*. Behavioural Brain Sciences Centre, University of Birmingham. https://doi.org/10.1017.S0142716405050034.

Bretz, F., Hothorn, T., & Westfall, P. (2010). *Multiple comparisons Using R*. London: Chapman and Hall/CRC. https://doi.org/10.1080/00401706.1964.10490181

Brouwer, A. M., Hogervorst, M. A., Holewijn, M., & van Erp, J. B. F. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *International Journal of Psychophysiology*, *93*, 242–52. https://doi.org/10.1016/j.ijpsycho.2014.05.004

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement* (Human– Computer Interaction Series). Cham: Springer International Publishing.

Retrieved from https://link.springer.com/book/10.1007%2F978-3-319-31700-7 https://link.springer.com/book/10.1007%2F978-3-319-31700-7

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, *25*, 315–24. https://doi.org/10.1016/j.chb.2008.12.020

Conway, D., Dick, I., Li, Z., Wang, Y., & Chen, F. (2013). The effect of stress on cognitive load measurement. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2013. INTERACT 2013. Lecture notes in computer science* (Vol. *8120*). Berlin, Heidelberg: Springer.

Cranford, K. N., Tiettmeyer, J. M., Chuprinko, B. C., Jordan, S., & Grove, N. P. (2014). Measuring load on working memory: The use of heart rate as a means of measuring chemistry students cognitive load. *Journal of Chemical Education*, *91*, 641–7. https://doi.org/10.1021/ed400576n

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*, 338–49. https://doi.org/10.1111/jcal.12288

Dindar, M., Malmberg, J., Järvelä, S., Haataja, E., & Kirschner, P. (2019). Matching self-reports with electrodermal activity data: Investigating temporal changes in self-regulated learning. *Education and Information Technologies*, *25*, 1785–1802. https://doi.org/10.1007/s10639-019-10059-5

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods*, *6*(3), 282–308. https://doi.org/10.1177/1094428103255532

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

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (pp. 301–310). Copenhagen. Retrieved from http://www.cs.cmu.edu/~sjunikim/publications/UBICOMP2010_Cognitive_Load.pdf

Herborn, K. A., Graves, J. L., Jerem, P., Evans, N. P., Nager, R., McCafferty, D. J., & McKeegan, D. E. F. (2015). Skin temperature reveals the intensity of acute stress. *Physiology and Behavior*, *1*, 225–230. https://doi.org/10.1016/j.physbeh.2015.09.032

Iani, C., Gopher, D., & Lavie, P. (2004). Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology*, *41*, 789–98. https://doi.org/10.1111/j.1469-8986.2004.00200.x

Kalyuga, S., & Singh, A. M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*, 831–852. https://doi.org/10.1007/s10648-015-9352-0

Karthikeyan, P., Murugappan, M., & Yaacob, S. (2012). Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *Journal of Physical Therapy Science*, *24*, 1341–1344. https://doi.org/10.1589/jpts.24.1341

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, *15*, 235–245. https://doi.org/10.30773/pi.2017.08.17.

Larmuseau, C., Vanneste, P., Cornelis, J., Desmet, P., & Depaepe, F. (2019). Combining physiological data and subjective measurements to investigate cognitive load during complex learning. *Frontline Learning Research*, *7*, 57–74. https://doi.org/10.14786/flr.v7i2.403.

Morton, J., Vanneste, P., Larmuseau, C., Van Acker, B., Raes, A., Bombeke, K., … De Marez, L. (2019). Identifying predictive EEG features for cognitive overload detection in assembly workers in Industry 4.0. In *Proceedings of the 3rd International Symposium on Human Mental Workload: Models and Applications (H-WORKLOAD)*. Rome, Italy. Retrieved from https://arrow.tudublin.ie/hwork19/1/

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Conference on Australian Computer-Human Interaction OzCHI '12*. Melbourne. https://doi.org/10.1145/2414536.2414602

Paas, F. (1992). Training strategies for attaining transfer of problem solving skills in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*, 429–34. https://doi.org/10.1037/0022-0663.84.4.429

Ramasubramanian, K., & Singh, A. (2016). *Machine learning using R*. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2334-5

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *Extended abstracts on Human factors in computing systems, 2651-2656. CHI'07*. San Jose, California. https://doi.org/10.1145/1240866.1241057https://doi.org/10.1145/1240866.1241057

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., … Van Hoof, C. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *Npj Digital Medicine*, *67*. https://doi.org/10.1038/s41746-018-0074-9

Spikol, D., & Cukurova, M. (2019). Multimodal learning analytics. In A. Tatnall (Ed.), *Encyclopedia of education and information technologies* (pp. 1–8). Cham: Springer.

Stone, J. M., & Towse, J. N. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*. https://doi.org/10.5334/jors.br

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *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*, 123–38. 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*, 261–92. https://doi.org/10.1007/s10648-019-09465-5

Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A., & Van Huffel, S. (2011). Instantaneous changes in heart rate regulation due to mental load in simulated office work. *European Journal of Applied Physiology*, *111*, 1497–505. https://doi.org/10.1007/s00421-010-1776-0

Van Merriënboer, J. J. G., & Sluijsmans, D. M. A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed learning. *Educational Psychology Review*, *21*, 55–66. https://doi.org/10.1007/s10648-008-9092-5

Yuan, K., Steedle, J., Shavelson, R., Alonzo, A., & Oppezzo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review*, *1*, 83–98. https://doi.org/10.1016/j.edurev.2006.08.005

Supporting Information

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