Cognition, Technology & Work (2021) 23:567–585
https://doi.org/10.1007/s10111-020-00641-0

ORIGINAL ARTICLE

Towards measuring cognitive load through multimodal physiological data

Pieter Vanneste[1,2] · Annelies Raes[1,2] · Jessica Morton[3] · Klaas Bombeke[3] · Bram B. Van Acker[3] · Charlotte Larmuseau[1,2] · Fien Depaepe[1,2] · Wim Van den Noortgate[1,2]

Received: 20 February 2020 / Accepted: 1 July 2020 / Published online: 12 July 2020

Abstract
Cognitive load plays an important role during learning and working, as it has been linked to well-functioning cognitive processes, performance, burnout and depression. Nonetheless, attempts to assess cognitive load in real-time by means of physiological data have been proven difficult, and interpreting these data remains challenging. The aim of this study is to examine whether and how well experienced cognitive load can be measured through psycho-physiological data. The approach of this study is rather unique, for a combination of reasons. First, this study takes a multimodal approach, monitoring EDA (electrodermal activity), EEG (electroencephalography) and EOG (electrooculography). Second, this study is based on a relatively intensive data collection ($N=46$) in a controlled lab setting in which varying cognitive load levels are deliberately induced. Finally, not only focussing on statistical significance but also on the size of the association gives insights into how suitable physiological markers are to measure cognitive load. Results from a multilevel analysis suggest that the following physiological markers might be related to cognitive load, for example, in an industrial context: the rate and the duration of skin conductance responses, the alpha power, the alpha peak frequency and the eye blink rate. About 22.8% of the variance in self-reported cognitive load can be explained using these five measures.

Keywords Cognitive load · Mental effort · Physiology · EEG · EOG · EDA

1 Introduction

1.1 The relevance of cognitive load in view of cognitive performance and well-being

Cognitive load has been studied in various scientific domains (i.e. cognitive psychology, instructional design, human factors and ergonomics), as it plays an important role in how well human cognitive processes function, as well as in psychological well-being.

The relationship between cognitive load and the functioning of cognitive processes follows from several studies (Chen et al. 2016; Johannsen 1979) arguing that too high levels of cognitive load (referred to as "cognitive overload"[1]) can have a negative impact on the performance of the working memory. Correspondingly, too low levels of cognitive load (referred to as "cognitive underload") can yield a worse performance as well, due to boredom or a lack of motivation (Young et al. 2014). These studies stress the importance of inducing adequate levels of cognitive load in view of well-functioning cognitive processes. As suboptimal cognitive load levels hamper cognitive processes, they may be detrimental to the productivity and quality of industrial processes (in a blue-collar context), or to the performance of office environment processes (in a white-collar context).

Cognitive load has also been studied in view of human–machine interaction applications. Wearable cognitive assistants that provide operators with the right

✉ Pieter Vanneste
pieter.vanneste@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[2] KU Leuven, Imec Research Group ITEC, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium

[3] Research Group Imec-Mict-UGent, De Krook, Miriam Makebaplein 1, 9000 Ghent, Belgium

[1] Note that the concepts of cognitive underload and overload may be somehow misleading, as the human working memory is limited: cognitive load can obviously not be inferior to the working memory's minimum capacity, nor can cognitive load exceed the working memory's maximum capacity. However, as these concepts are intuitively easy to understand, we will refer to them sometimes.

information at the right time offer a way to minimize errors, to enhance the person job-fit and to keep the cognitive load in the "comfort zone" (Belletier et al. 2019).

Next to cognitive load's impact on the functioning of cognitive processes and on performance, the cognitive load has also been related to psychological well-being. A study from Iskander (2018) among medical professionals argues that cognitive overload is likely to be an immediate precedent of burnout. The author stresses the importance of metacognitive training to monitor one's own cognitive load as a skill to help prevent burnout. Furthermore, as it is mostly agreed upon that given the plurality of depressive disorders, burnout and depression are connected conditions (Schonfeld and Bianchi 2016), tracking cognitive (over)load could also be highly relevant for the prevention of depression. More evidence hereto can be found in the systematic review from Bianchi et al. (2015).

Given the relevance of cognitive load in terms of cognitive performance and well-being, keeping track of cognitive load is, therefore, an important endeavour in fundamental research, but may also have interesting applications in practice. Although this study can be understood in a relatively wide range of applications, a specific context is highlighted as an example: industrial assembly work.

In industry, the cognitive load of operators during assembly tasks is likely to depend on the complexity of the task at hand, the work instructions that are provided and the experience of the operator. Industry 4.0 refers to the fourth industrial revolution, following the previous mechanical, electrical and digital revolution. In this fourth industrial revolution, an advanced digitalization transforms factories in smart organizational units in which technological components are increasingly interconnected via sensors and via the internet (Lasi et al. 2014). Smaller lot sizes and an increasing product diversification are another typical property of Industry 4.0. The range of products that manufacturing companies deliver (i.e. product-portfolios) becomes even more diversified and new products are being introduced more rapidly. This leads to increasingly complex manufacturing processes (El Maraghy et al. 2012; Wan and Sanders 2017).

In this increasingly complex industrial ecosystem, operators spend less time on repetitive processes, as these can increasingly be fulfilled by machines. This also means that operators can rely less on routine skills. They are confronted with new learning tasks on a more frequent basis, which is, from a cognitive perspective, a challenging evolution. Diversified and rapid manufacturing may result in more information processing (understanding how to assemble increasingly complex products) and information storage (the need to remember for more products how they need to assembled). Note that this effect of complexity on cognitive load is not a priori given, as it depends on the application at hand. As an example, the increased automation and interconnectivity of Industry 4.0 may, to a certain extent, also facilitate these complex processes for the operator.

Technological evolutions have opened the possibility to keep track of numerous types of data, such as physiological data, which can indirectly reflect humans' emotions. However, it remains unclear how and to what extent cognitive load is reflected in these data. Noroozi et al. (2019) state that it is not yet clear which constructs (self-regulation, cognitive load, attention, engagement, etc.) different physiological signals can measure and how the data should be triangulated and interpreted in the light of the construct that is being studied. Brouwer et al. (2015) mention that the potential of neurophysiological signals to infer mental states is often overestimated, mainly because conclusions are not always warranted and generalizations made are potentially problematic. The authors attribute these deficiencies to two main root causes: the highly interdisciplinary nature of the field which makes it difficult to master all aspects on the one hand, and the unjustifiable belief to consider neurophysiological data as conveying an objective truth on the other hand. To overcome these deficiencies, the authors give six recommendations to avoid common pitfalls when performing research in this scientific domain: defining a ground truth, formulating hypotheses about the link between neurophysiological measures and the mental state of interest, eliminating confounding factors, using proper statistical analyses, providing insight into the data and clarifying the added value of using neurophysiology.

To contribute to clarifying how experienced cognitive load is reflected in physiological data, this study collects multimodal physiological data in a controlled lab environment, in which various levels of cognitive load are induced. In doing so, this study aims to respond to the six aforementioned recommendations of Brouwer et al. (2015).

In what follows, first the concept of cognitive load is defined. Next, traditional ways of assessing experienced cognitive load by means of self-reporting are described. Hereafter, physiological measures are discussed, and the theoretical underpinning for their relatedness with cognitive load is explained. In addition, the main findings from studies that also aim to measure cognitive load by means of physiological data are listed. This is followed by the research aim and a detailed description of the employed methodology. Finally, the results are reported and discussed.

1.2 Definition of cognitive load

The concept of cognitive load originates from early work in the field of instruction and education and is specified in the so-called Cognitive Load Theory (CLT; Sweller 1988; Sweller 1994; Sweller et al. 1998). CLT defines cognitive load as the demands being put on the storage and the processing of information in the human working memory

(Schnotz and Kürschner 2007). In cognitive psychology, cognitive load is defined similarly, as the amount of working memory resources that is used (Chen et al. 2016). Cognitive load is also studied in ergonomics and human factors literature, where it is referred to as mental workload, mental effort or mental demand and defined as the amount of mental activity required to perform a task (e.g., Van Acker et al. 2018; Young et al. 2014). In sum, definitions of cognitive load differ from domain to domain but show a clear common ground, i.e. the proportion of the human working memory capacity that is addressed. This capacity is limited, which is in contrast to our sensory and long-term memory that are able to process (and store) a quasi-unlimited amount of information.

1.3 Traditional self-reporting assessment scales for cognitive load

A traditional and relatively simple way to measure (experienced) cognitive load is via self-reports, which are often considered as a gold standard. Previous research has used both unidimensional and multidimensional self-reporting scales. Multidimensional scales distinguish several components out of which cognitive load consists. These components typically depend on the way in which cognitive load is conceptualised, and can thus differ across scientific domains. In instructional design research, for example, cognitive load encompasses three constructs: the cognitive load associated with interpreting the learning instructions, understanding the actual content, and storing the acquired knowledge (Leppink et al. 2013). In the field of ergonomics and human factors, Reid and Nygren (1988) address mental workload by means of their Subjective Workload Assessment Technique (SWAT) and state it can be largely explained by three components: time load, mental effort load and physiological stress load. Although mental workload definitions are sometimes divergent, cognitive load is often closely linked to the mental effort load component (Young et al. 2014). Another commonly used multidimensional scale in the field of ergonomics and human factors is the NASA Task Load Index (NASA-TLX, Hart and Staveland 1988) that assesses workload on five scales with 21 gradations each. The mental demand component of the NASA-TLX closely aligns to cognitive load. Next to multidimensional scales, unidimensional scales have also been proven to be reliable and valid assessment tools of cognitive load. A frequently used unidimensional scale in cognitive load research is Paas' nine-point mental effort rating scale (Paas 1992). The scale ranges from "*very, very low mental effort*" to "*very, very high mental effort*" and assumes that learners can retrospectively assess their own cognitive load.

Self-reports are widely used and are considered a valuable source of information by a large body of literature. Fryer and Dinsmore (2020) for instance claim that in certain instances, self-reporting may be the only viable way to unearth covert constructions, such as emotional or cognitive states. Pekrun (2020) acknowledges certain limitations of self-reporting but argues they are still indispensable for any more nuanced assessment of mental states. Although the construct validity of self-reporting can be disputed, self-reports at least directly inquire the construct of interest. For these and other more practical reasons, subjective measures have been used in research for decades as a valuable source of information and sometimes tend to be seen as a "gold standard".

However, self-reports also have several disadvantages. First, they do not permit measuring a person too frequently, let alone continuously (in real-time) (Matthews et al. 2019; Young et al. 2014). In addition, self-reports are intrusive, as they interrupt subjects by redirection their attention from the mental state they are into the self-report measure (Zimmerman 2008). Also retrospective self-reporting has disadvantages, for instance because the self-report may be a post-hoc reconstruction rather than a real reflection of the actual cognitive load. Finally, it is important to be aware of other types of biases that self-reports are prone to, such as individual differences in interpreting the question and rating the numerical scales. Note that the existence of several types of biases that self-reports are prone to implies that the observed associations between cognitive load and its potential indicators may be underestimating the real associations.

Sensor data are not prone to these shortcomings. These (objective) data allow automatic and real-time measurements without subjects' involvement and thus remedy the aforementioned shortcomings of self-reporting data. Although one may consider certain physiological measurement devices (EEG headsets, wristbands, patches, etc.) as obtrusive as well, it is expected that technological developments will lead to comfortable and wearable sensors in the future. Indeed, Zheng et al. (2014) give an overview of emerging unobtrusive wearable technologies, and explain how technological evolutions (micro- and nanotechnologies, mobile communications, human computer interfaces, etc.) are continuously making wearable sensors less obtrusive. The authors give examples of sensors that can be weaved or integrated into clothing, accessories and even the human skin. They explain how these developments enable to acquire information from these sensors in less interfering ways.

Understanding how the self-reported cognitive load relates to physiological data would enable to measure cognitive load in real-time and in a non-obtrusive way (without interrupting the subject), through directly measurable manifest variables. The (construct) validity of physiological data in terms of measuring cognitive load is obviously crucial, and depends on the extent to which cognitive load induces a certain physiological reaction. The next paragraphs discuss this in more detail.

1.4 Psycho-physiological measures for cognitive load

Previous research has shown that there is a theoretical ground for a link between cognitive load and physiology (Chen et al. 2016; Haapalainen et al. 2010; Kramer 1990). Several types of physiological measures have already been addressed in the context of cognitive load measurement, such as electrodermal activity (EDA), skin temperature, electrocardiograms (ECG, reflecting different heart rate measures, including heart rate and heart rate variability), electroencephalography (EEG) including event-related potentials (ERPs), electrooculography (EOG), functional near-infrared (fNIR) spectroscopy (e.g., Ayaz et al. 2012; Liu et al. 2017) and eye tracking (e.g., pupillometry).

Table 1 lists the different physiological measures that have often been used in an aim to measure cognitive load. First, a short description of the physiological measure itself is given. Next to that, an elaboration is given of the neurobiological underpinning on which the presumed relationship between the physiological measure and cognitive load relies. For some physiological measures, this table suggests a relatively direct association with cognitive load. This is the case for EEG measures (both the power of frequency bands and event-related potentials) as well as for EOG measures, eye tracking and pupillometry. This rather narrow neurobiological link opens up the possibility for strong associations between these physiological measures and cognitive load. However, for EDA measures, skin temperature and heart rate measures, the neurobiological link is much more indirect, which suggests that their association with cognitive load may be weaker, or potentially non-existing.

Kramer's overview (1990) on physiological measures for mental workload stresses that given the inherent multidimensional nature of cognitive load, no single measurement technique can capture all its aspects. Each physiological measure will perform differently related to the five main criteria that Kramer mentions: sensitivity, diagnosticity, intrusiveness, reliability and generality of application. In addition, it is acknowledged that multimodal approaches can overcome the limitations of single-source measurements and provide more robust representations of cognitive load (Chen et al. 2016).

Whereas Table 1 is primarily theoretically oriented, Table 2 dives into findings from previous empirical studies. Table 2 gives a non-exhaustive overview of the characteristics and findings of different studies that have investigated how the cognitive load is reflected in physiological measures.

Several limitations can be observed from the studies listed in Table 2. These limitations are directly linked to the research gaps that are mentioned in Sect. 2, Research aim.

Another important deduction of Table 2 is that for some physiological measures no significant associations with cognitive load are observed. This is the case for EDA measures, skin temperature and heart rate. For other measures, there is evidence that they are related to cognitive load. This applies to heart rate variability as well as to EEG (mainly the alpha activity). Finally, concerning EOG, results suggest that several eye measures are associated with cognitive load: the blink rate (and interval), the blink latency, the pupil microsaccades magnitude and the pupil diameter. We can also conclude from the literature review that the strength of the associations (effect sizes) with cognitive load is not always mentioned, and when it is mentioned, effect sizes are rather moderate or small.

When measuring latent variables by means of (a combination of) manifest variables, we are not merely looking for significant relationships between the manifest variables and (a proxy of) the latent variable, but especially for manifest variables that are strongly related to the latent variable, and, therefore, can be considered as reliable and valid indicators.

Haapalainen et al. (2010) for example found two significant physiological measures to discriminate cognitive load (see Table 2) that could predict the complexity condition of the task (as either high or low) with an accuracy of 81%. Important to mention is that this accuracy is an average and results from individual models that were created for each participant separately. The authors also attempted to find a single model across all participants to discriminate the different complexity levels, but mention that due to individual differences between participants they have not yet been able to do so. This shows how challenging it is to measure cognitive load at the individual subject's level.

Fisher et al. (2018) express their concern about research that unjustifiably applies "group-to-individual" generalizability, and argue how this results in imprecise and potentially invalid conclusions. They explain that only for ergodic processes, inferences based on associations across individuals also generalise to the individual level. Typical of such processes is that the mean and the variance of the construct of study do not vary over time, which is rarely the case in research that studies human behaviour. The authors evaluated six studies with a repeated measures design and found that the variance within individuals was two to four times larger than the variance between individuals. The authors state that "the highest-impact publications in medical and social sciences have been largely based on data aggregated across large samples, with best-practice guidelines almost exclusively based on statistical inferences from group designs" (p. 1).

Table 1 An overview of most commonly studied physiological measures and a brief theoretical underpinning for their link with cognitive load

Physiological measure	Brief description and theoretical underpinning of the presumed association between the physiological measure and cognitive load
EEG: power of frequency bands	*Brief description:* EEG allows to measure brain activity rather noninvasively. Performing a spectral analysis on the measured electric potential differences (by means of a fast-Fourier transform, FFT) allows to analyse the power of different frequency bands (delta, theta, alpha, beta and gamma) that are present in the signal
	Underlying hypothesis: An increase in cognitive load can be measured by an increase in brain activity, i.e. oscillations within a certain frequency band with a larger amplitude (Antonenko et al. 2010)
EEG: Event-related potentials	*Brief description:* An example of an event-related potentials approach is to let humans listen to beep tones, of which a small fraction has a deviating frequency (i.e., some beeps are higher or lower in tone)
	Underlying hypothesis: Humans, whether conscious or not, process auditory stimuli via their sensory working memory. This is reflected in brain activity, which can be measured by a voltage difference before and after the beep tone. Previous research has found that higher cognitive load levels result in less profound working memory processing and thus in smaller voltage differences (E.g. Luck 2012)
EOG	*Brief description:* EOG (i.e., electrooculography, deploying the same sensors as EEG) enables to assess eye blinks in a non-obtrusive way
	Underlying hypothesis: During high cognitive load levels, eye blinking is reduced, as blinking interrupts visual information, which is then undesirable (e.g. Ledger 2013). Next to an increase in endogenous eye blink rate (other than reflexive and voluntary eye-blinks), a decline in blink closure time and an increase in blink latency (the time between a stimulus and the blink initiation) have been associated with higher cognitive load levels as well (e.g., Zagermann et al. 2018)
Eye-tracking and pupillometry	*Brief description:* Measuring the pupil diameter, blink latency, eye fixations and saccade characteristics
	Underlying hypothesis: Eye-tracking and pupillometry have been related to cognitive load in previous studies (e.g. an increase in cognitive load has been associated with pupil dilations) via neurobiological mechanisms underlying the link between these eye-measures and cognitive load (e.g., Van der Wel & Van Steenbergen 2018), such as the innervation of neurons of the autonomic nervous system with radial fibres of the iris
EDA	*Brief description:* Electrodermal activity (EDA), which is also referred to as Galvanic Skin Response (GSR) assesses electrical characteristics of the skin to infer changes from the sympathetic nervous system
	Underlying hypothesis: A psychological state that causes a human to experience stress or arousal, will give rise to an increase in skin conductance. The hypothesis for the cognitive load to be reflected in EDA is that cognitive load has an effect on stress or arousal, which in turn impacts EDA (e.g. Setz et al. 2010)
Skin temperature	*Brief description:* Measuring the temperature of the outermost surface of the human body
	Underlying hypothesis: Stress or arousal results in vasoconstriction (narrowing of the blood vessels from the human skin), which reduces the temperature of the skin. The hypothesis for the cognitive load to be reflected in the human skin temperature implies that cognitive load has an effect on stress or arousal, which in turn impacts the skin temperature (e.g. Herborn et al. 2015)
Heart rate measures	*Brief description:* Heart rate and heart rate variability can be assed via electrocardiograms or photoplethysmography (a light-based technology)
	Underlying hypothesis: Heart rate is directly linked to physical activity, but is also a measure of both the sympathetic and parasympathetic autonomic nervous system activity. However, these only have a very indirect link with cognitive load (Jerčić et al. 2018). Heart rate variability is often used as an indication of the modulation of the autonomic nervous system. Stress or arousal will cause an increase in blood pressure and a decrease in heart rate variability (Solhjoo et al. 2019), which makes the link with cognitive load also indirect

2 Research aim

A literature review shows that there are several studies that have explored and evaluated the significance of a psychophysiological measure in the light of studying cognitive load. However, several research gaps can be identified.

A first research gap is that not all studies take an advanced multimodal approach: some only address a single or a few physiological markers, which may prevent capturing all aspects of cognitive load.

A second research gap concerns different aspects that relate to the design and the methodology of previous studies. A first aspect is that most studies involve multiple complexity conditions, but do not deliberately inquire about the induced cognitive load. Nonetheless, cognitive load is a subjective experience and not only the effect of the context on cognitive load but also the way cognitive load is manifested can be person-dependent. In this respect, self-reports are well suited to capture each participant's subjective experience for each condition. A second aspect is that existing studies do typically not include a high complexity condition, intended to induce cognitive overload, although such a condition can make associations between cognitive load and manifest variables more

Table 2 A non-exhaustive overview of different studies that have investigated how (an increase in) cognitive load is reflected in physiological measures

Study	Analysis technique	Sample size	EEG (power of frequency bands, event-related potentials)	Eye measures: EOG, eye-tracking and pupillometry	EDA, skin temperature and heart rate measures	Indication of the effect size or the proportion of explained variance when all investigated measures are combined in one model
Ryu and Myung (2005)	Multiple regression model	$N=10$	Alpha activity suppression ($p<0.05$)	Blink interval: increase ($p<0.01$)	Heart rate: increase ($p<0.05$)	$R^2_{adj}=0.51$
Haapalainen et al. (2010)	Machine learning: Naive Bayes classifier	$N=20$	Delta, theta, alpha, beta and gamma power: N.S	Pupil diameter: N.S	–Skin conductance: N.S. –Skin temperature: N.S. –Heart rate: N.S.; heart rate variability: N.S.	Prediction accuracies of composite scores do not surpass chance level (significances were only observed for individual models per participant, see further)
Antonenko et al. (2010)	Review of literature on the use of EEG to measure cognitive load		–Alpha activity suppression (desynchronization at parietal regions) –Theta activity increase	–	–	The literature review focuses on the use of EEG. A model combining multiple physiological measures is not addressed
Marquart et al. (2015) (literature review)	Review of the literature, analysing studies that address eye measures to reflect changes in cognitive load (context: driving)		–	–Pupil dilation –Blink latency increase –Eye fixation duration increase –Blink rate: mixed results	–	A model combining multiple physiological measures is not addressed
Krejtz, et al. (2018)	ANCOVA and multinomial logistic regression	$N=13$	-	–Pupil micro-saccades: greater magnitude ($p<0.001$, $\eta^2=0.17$), but no differences in rate nor peak velocity ($p>0.05$); –Higher intra-trial change of pupil diameter ($p<0.05$, $\eta^2=0.07$); –Higher inter-trial change in pupil diameter ($p<0.001$, $\eta^2=0.16$).	–	–Despite the highly significant findings for different eye measures, effect sizes only allow to explain a limited proportion of variance in cognitive load –Significant increases in the log odds of a certain complexity condition are found when multiple physiological measures are combined. Effect sizes are not considered
Rosch and Vogel- Walcutt (2013)	Review of the literature, analysing studies that address eye measures to reflect changes in cognitive load		-	Longer eye-gaze fixation durations, higher saccade peak velocities and more saccade errors. Significance levels or effect sizes are not mentioned	-	A model combining multiple physiological measures is not addressed

Table 2 (continued)

Study	Analysis technique	Sample size	EEG (power of frequency bands, event-related potentials)	Eye measures: EOG, eye-tracking and pupillometry	EDA, skin temperature and heart rate measures	Indication of the effect size or the proportion of explained variance when all investigated measures are combined in one model
Larmuseau et al. (2019)	Multilevel model	$N=15$	–	–	–Skin conductance: N.S. Only significant differences (both $p<0.05$) between the baseline measurement and the low (Cohen's $d=0.19$) and high complex (Cohen's $d=0.14$) conditions – Skin temperature: N.S	A model combining multiple physiological measures is not addressed
Cranford et al. (2014)	Friedman test, a non-parametric test equivalent to repeated measures ANOVA	$N=19$	–	–	Heart rate: when compared to a baseline measurement, the average increase was higher for high complexity conditions than for low complexity conditions ($p<0.05$)	–Effect sizes are not mentioned –Despite the significant findings across individuals, there is still a large between-subjects variation, as the standard deviation in heart rate amongst participants is larger than most differences in mean heart rate between the conditions themselves
Solhjoo et al. (2019)	Correlation analysis	$N=10$	–	–	–Heart rate: N.S –Heart rate variability: increase ($p<0.05$, $R^2=0.50$)	$R^2=0.28$–0.81

– not investigated, *N.S.* not significant; the significance (*p*-value) is indicated unless it is not mentioned by the study.

visible. A third aspect is that quite some studies lack statistical power, because of a rather limited sample size and because each participant is only once or a few times measured within the same condition.

A third research gap relates to the statistical analyses and the interpretation of the results in view of implications for research and for practice. A first aspect is that if studies include repeated measurements, these are sometimes not statistically analysed in an appropriate way. A second aspect is that often no measure is mentioned of how well manifest variables succeed in measuring cognitive load (such as the proportion of explained variance), or if such a measure is reported, it is not interpreted in the light of using the studied physiological markers as a measurement tool for the cognitive load. Nonetheless, such a measure of association or goodness of fit represents an important criterion if one wants to actually consider the use of physiological data as a measurement instrument for the cognitive load.

The aim of this study is to simultaneously meet these research gaps, and to examine whether and how well participants' experienced cognitive load can be measured through psycho-physiological data.

To effectively measure cognitive load, a measurement model is required that takes several manifest variables as input and is thereby capable to measure cognitive load sufficiently precise and in an automatic way. However, the exact appearance of such a measurement model is not self-evident. This study monitors EEG, EOG and EDA data, as previous studies have shown that these physiological sources might be promising in terms of measuring cognitive load.

To be more precise, this study pursues the following objectives:

- Investigate how well we can measure the latent experienced cognitive load, using a stringent methodological approach, by means of the following physiological manifest variables:
  - EDA and skin temperature,
  - EEG,
  - EOG, and
  - a composite score based on EDA, skin temperature, EEG and EOG.
- Uncover the possibilities and limitations of measuring cognitive load through physiological data to evaluate the corresponding implications both for research and for practice.

3 Methodology

3.1 Participants

Participants in this study were recruited in January 2019 in Gent (Belgium), in a public library which is situated in the same building as a university (see also Morton et al. 2019). In total, 46 participants voluntarily signed up for the study and received a small financial reward. They were aged between 19 and 40 years old (the average age was 25.8 years, as most participants were students). There were 25 female and 21 male participants.

3.2 Experimental design and tasks

To manipulate cognitive load across the different conditions, we act on two processes from the working memory, namely storing (remembering) information and processing information (Sweller 2010). According to Kyllonen and Christal (1990), good measures for working memory should (1) include simultaneous processing and storage, (2) not involve learning and (3) require knowledge that all subjects are presumed to have.

A first method to vary cognitive load is by manipulating (visuo-spatial) information processing, through the difficulty of the task. For that purpose, tangram puzzles are used. These are dissection puzzles that consist of seven flat pieces with different sizes and forms. These individual pieces have to be arranged together in a certain way and without overlap to form a shape. As the pieces can be put together in a quasi-unlimited number of ways, many different shapes can be formed. The difficulty of the puzzle stems from the extent to which only contours (outlines) are shown, which masks the way in which the individual pieces should be arranged to form the required shape. In the low complexity phase, participants assembled several tangram puzzles of which the contours of each of the seven pieces are individually visible. In the medium complex phase, all puzzles have three pairs of two pieces touching each other, so only their surrounding contour is visible, which requires more (mainly visuo-spatial) information processing to find out how they should be assembled. In the highly complex phase, all puzzles have multiple touching sides, which makes it even more difficult to find out how the puzzle should be assembled (this is also illustrated in Fig. 1). This method to manipulate information processing through the difficulty of the task fits in the framework proposed by Richardson et al. (2006) in which several variables that predict object assembly difficulty are identified. Applied to Tangram puzzles, a higher complexity is characterised by more possible ways to orient and align the pieces as well as a higher amount of symmetrical planes.

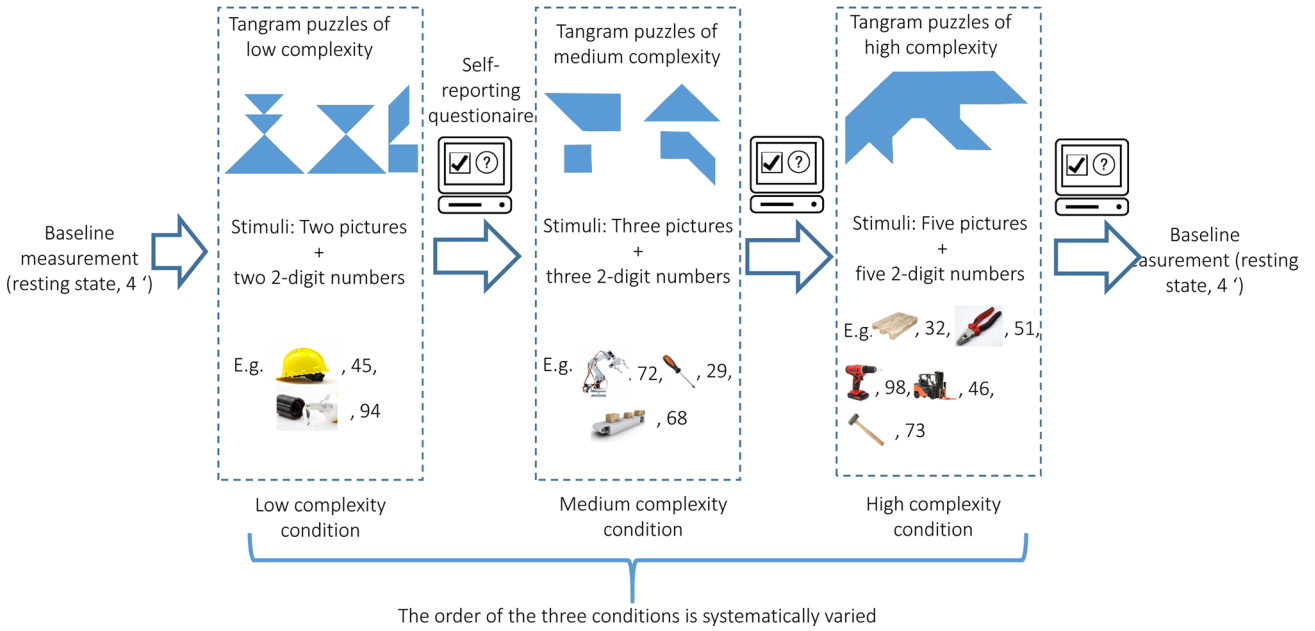

Fig. 1 The design and procedure of the study

A second method to vary cognitive load is by manipulating information storage, by varying the type and amount of visual stimuli. Each stimulus is shown for a duration of 30 s on a computer screen in front of the participant while performing the tangram tasks. The time intervals between the different stimuli are held constant. The participants were asked to remember the stimuli and write down the ones they remember after each phase. The number of stimuli that is shown increased with the increasing complexity of the condition. Two different kinds of stimuli were alternately shown: pictures representing a tool that is typically used in industry (such as a safety helmet, a conveyor belt or a drilling machine) and a two-digit number. During the low complex phase, two pictures and two numbers were shown. During the medium complex phase, three pictures and three numbers were shown. During the high complex phase, five pictures and five numbers were shown.

3.3 Procedure

All participants were exposed to the procedure illustrated in Fig. 1. At the beginning of the data collection, a baseline measurement was performed in which the participant was in a quiet condition (resting state) and no tasks needed to be performed. During that baseline measurement, the participant first had to close his/her eyes for two minutes, after which (s)he had to look ahead with his/her eyes open for two minutes. During the entire baseline phase, EEG, EOG and EDA data were collected. Collecting baseline measurements enables to account for the highly individual nature of physiological data, by comparing each participant's data collected during the different conditions to their own baseline values (see 3.5.1., Data pre-processing).

Subsequently, the participant went through three phases of ten minutes each, which are characterized by a different complexity level: a low complexity, a medium complexity or a high complexity. In doing so, we aim to induce three levels of cognitive load: a low level of cognitive load, a medium level of cognitive load and a high level of cognitive load (or cognitive overload, as it was intended to approximate the participants' maximum cognitive load). For each condition and for each participant, data were collected during a time span of 10 min. Because the time to complete one puzzle typically ranges from less than a minute to a few minutes, sufficient puzzles of the same complexity were made available to span the ten minutes time period. As such, during each condition, participants can assemble as many puzzles as they can, until the ten minutes period ends. By increasing the variation in cognitive load, we aim to ease the assessment of the value of physiological measures as indicators. To avoid learning or order effects, the sequence of the tasks was varied over participants by applying counterbalancing. This was operationalized by systematically alternating all six possible task sequences across participants.

Self-reported data inquiring the perceived cognitive load were collected three times, each time after the completion of a condition (see next subsection).

After measuring the participants under the three conditions, another baseline measurement was performed in which the participant is again in a resting state.

3.4 Apparatus to measure the physiological data and the self-reported data

The following physiological data (manifest variables) are monitored (each of the physiological measures is aggregated per participant over the entire ten minutes length of the condition):

- Measured with a Biosemi ActiveTwo (BioSemi, Amsterdam, Netherlands):
  - EEG data (with a focus on the power of the alpha frequency band and the maximum frequency within the alpha band)
  - EEG Event-Related Potentials: The N200 voltage difference (a usually negative voltage difference assessed 200 ms after the initiation of the beep tone)
  - EOG data: eye blink rate (via external electrodes, horizontally and vertically relative to the pupil)
- Measured with the imec Chillband + (imec, Leuven, Belgium):
  - EDA measures: tonic component of skin conductance, phasic component of skin conductance, rate of skin conductance responses, duration and magnitude of these skin conductance responses.
  - Skin temperature
  - Acceleration of the participants' left wrist (an indication for movement intensity)
  - Heart rate measures are monitored, but could not be included in the analyses, as the calculation algorithm to derive heart rate measures from photoplethysmography (light-based technology) was not reliable enough.

The interpretation of the different EDA measures deserves some additional explanation. The tonic skin conductance can be understood as the component of the skin conductance that changes slowly over time and is not impacted by sudden stimuli. The phasic skin conductance, on the other hand, shows up as abrupt and short-term increases in the skin conductance signal, which are caused by external stimuli, typically related to stress or arousal. The skin conductance response rate is a measure for the frequency in time at which such phasic peaks occur. These phasic peaks are short, but can still differ in duration, a phenomenon which is characterised by the skin conductance response duration. Finally, the skin conductance magnitude is the integral of the phasic skin conductance over time and is as such related to both the duration and the magnitude of the phasic peaks.

The latent variable, cognitive load, is retrospectively assessed after each condition. A similar unidimensional approach is used as Paas (1992), but instead of using a 9 point rating scale, this study inquires cognitive load digitally via a quasi-continuous scale, on which participants can indicate scores ranging from 0 to 100 by means of a slider. We consider this subjective self-report as a gold standard for experienced cognitive load and use it as a criterion to find suitable physiological indicators.

- In addition, also the following variables are kept track of in each condition:

  The perceived complexity: to be rated by participants on a 7 point Likert scale. The purpose of including this question is to assess whether our manipulation is successful, i.e. if the different conditions indeed induced different levels of perceived complexity (with an aim to consequently induce different levels of cognitive load).
- The number of correctly assembled tangram puzzles: a first performance indicator (mainly linked to processing information)
- The proportion of correctly remembered stimuli: a second performance indicator (mainly linked to remembering information)

3.5 Data analysis

3.5.1 Data pre-processing

Prior to the actual analysis, the data are first pre-processed. Spectral analysis is performed on the EEG data, transferring the time domain to the frequency domain, via a fast Fourier transform (FFT). Hereafter, a baseline correction is applied, correcting the actual measured value X (raw data) for the baseline measurement B. This is a common practice for physiological data, as they are highly person-dependent. For EEG data, a decibel baseline correction is applied, $10*10\log(X/B)$. For the EDA data, absolute baselining is applied, deducting the baseline measurement from the actual measurement ($X-B$). EOG is operationalised by means of the blink rate, which is obtained by dividing the total amount of detected blinks during a condition by the duration of that condition.

3.5.2 Manipulation check of the perceived complexity experienced cognitive load and performance

To check the manipulation of the perceived complexity and cognitive load, we compare these scores in the three conditions by means of descriptive side-by-side boxplots. In addition, multilevel analyses are conducted in which we

use the complexity condition as a categorical predictor, and perceived complexity or self-reported cognitive load as a criterion variable. Subjects' effects are included as random effects. In this way, we take into consideration that the within-subjects residuals are not independent: because the physiological measures (and the effect of the conditions on these measures) are likely to be person-dependent, deviations of the observed scores from the same person are likely to be more similar than deviations from different persons. Acknowledging this dependency is important, as failing to do this could possibly lead to flawed standard errors and therefore unjustified significant associations. Pairwise comparisons between conditions are performed, in which *p*-values are corrected for multiple testing according to Holm's method.

Possible order effects are also investigated, by including an interaction effect between the condition (low, medium or high complexity) and the order in which the participant was exhibited to that condition (first, second or third).

Next, the performance measures are plotted. They are analysed by means of a multilevel approach in which the dependent variables are the proportion of remembered stimuli and the number of correctly assembled puzzles. The complexity condition is taken into account as a fixed effect and the subject as a random effect.

3.5.3 Evaluation of the physiological measures as indicators

After the pre-processing of the data (as described earlier), the data are explored by means of correlation matrices. These correlation matrices give a first idea about how the physiological measures interrelate with each other and with the self-reported cognitive load. These analyses are not conclusive, as they do not account for repeated measures (for each variable, we have three scores per participant, i.e., one score per condition) nor for possible confounding effects of other predictor variables.

Next, the data are analysed via a multilevel approach, regressing the self-reported cognitive load on the physiological data. By allowing the intercept to vary among subjects, we explicitly model that self-reported scores from the same participant can be systematically low (or high).

Multilevel (or mixed effects) models are well suited to address the "group-to-individual" generalizability concern as they allow to assess the proportion of variance in self-reported cognitive load that the predictor variables can explain. More specifically, we are interested in describing the within-subjects or residual variance (rather than the between-subjects variance) using the physiological indicators. By comparing the proportion of additional explained residual variance between a null model without predictors and another model that does include predictor variables, we can assess the proportion of residual variance that these predictors can explain.

Another advantage of multilevel models is that they can disentangle between-subjects variation (which arises from inter-individual differences, such as participants systematically scoring higher than others, which is very common for physiological measures) from within-subjects variation (e.g., arising from actual differences in perceived cognitive load between conditions). Less sophisticated statistical techniques such as bivariate correlations, for example, are often used in research, but cannot make this distinction. As a result, these simple correlation measures can represent an underestimation of the true relationship between self-reported cognitive load and a certain physiological measure.

A final advantage is that multilevel models are insightful: the relationship between the dependent variable and the predictor variables follows directly from the obtained model.

All physiological measures are centered around their mean, so the intercept of the different models refers to the expected value for the self-reported cognitive load score (outcome variable *Y*) when all physiological measures are equal to their mean value. The regression coefficients β express to what degree the latent variable of interest, the experienced cognitive load, is expected to increase with one-unit increases of the potential physiological indicators.

In the first step, each physiological feature is included in a separate multilevel model. In the second step, the most distinct measures are included together in three multilevel models. Measures for which the Variance Inflation Factor (VIF) is larger than 10 are excluded from the analyses, given their high multicollinearity. Models are made respectively for the physiological data measured with the imec Chillband+, for EEG data and for EOG data. These models give an idea about how well these types of physiological data can indicate cognitive load.

Finally, the physiological features that are most explanatory for cognitive load, regardless of their "type", are included in a single model. This model will elucidate how well cognitive load can be measured when combining the best indicators across all "types" of physiological data (EDA, skin temperature, EEG and EOG).

4 Results

4.1 Manipulation check of the perceived complexity experienced cognitive load and performance

Results indicate that participants perceived the study's three conditions indeed different in terms of complexity, $F(2,90) = 263.9$, $p < 0.001$, $R^2 = 0.80$. This is also illustrated by Fig. 2 (left side). In addition, pairwise comparisons show that each condition differs from each other condition

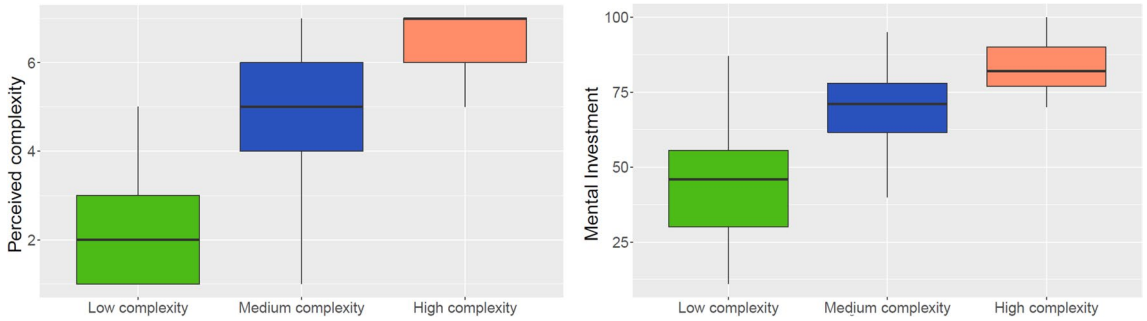

Fig. 2 Participants' perceived complexity (figure on the left) and mental investment (figure on the right) across the different conditions

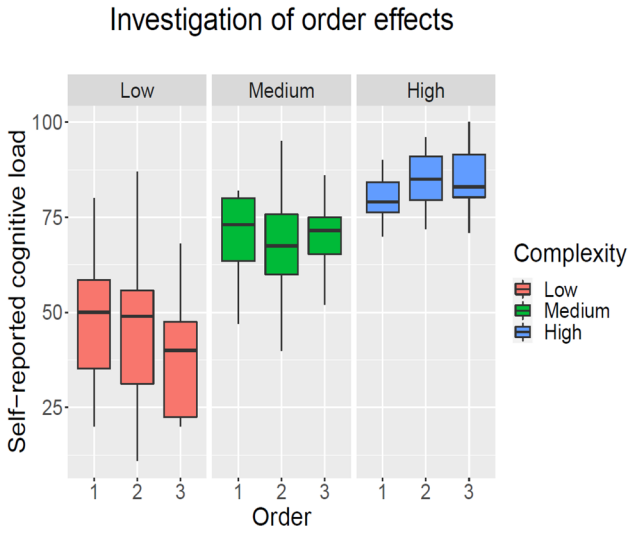

Fig. 3 Distribution of participants' self-reported cognitive load scores (y-axis). Within blocks representing the levels of complexity, the x-axis indicates the order in which a participant is subjected to a certain complexity condition

in terms of perceived complexity (all $p < 0.001$). Moreover, these complexity levels induced three different levels of cognitive load, $F(2,84) = 117.3$, $p < 0.001$, $R^2 = 0.66$, as can be seen in Fig. 2 (right side), supported by the results from pairwise comparisons (all $p < 0.001$).

In sum, conditions that required more information processing and storage were perceived as more complex, and induced a higher cognitive load. These findings are indications that we were able to manipulate the (experienced) cognitive load, which will make it easier to answer the research questions.

An investigation of order effects (see Fig. 3) reveals that when the low complexity task is performed as the last of the three phases, it induces (on top of the general effect that the low complexity has on cognitive load) a lower cognitive load ($p = 0.01$, $\beta = -16.1$, S.E. $= 7.6$).

Figure 4 displays the performance measures across conditions. An interesting observation is that the proportion of remembered stimuli decreases across conditions ($F(2,90) = 51.7$, $p < 0.001$, $R^2 = 0.52$, and for all pairwise comparisons, $p < 0.001$). This means that participants tend to remember a smaller proportion of pictures as the total number of pictures that is shown increases. In addition, the number of tangram puzzles that participants assembled correctly also decreases with increasing complexity ($F(2,90) = 539.1$, $p < 0.001$, $R^2 = 0.89$ and for all pairwise comparisons, $p < 0.001$).

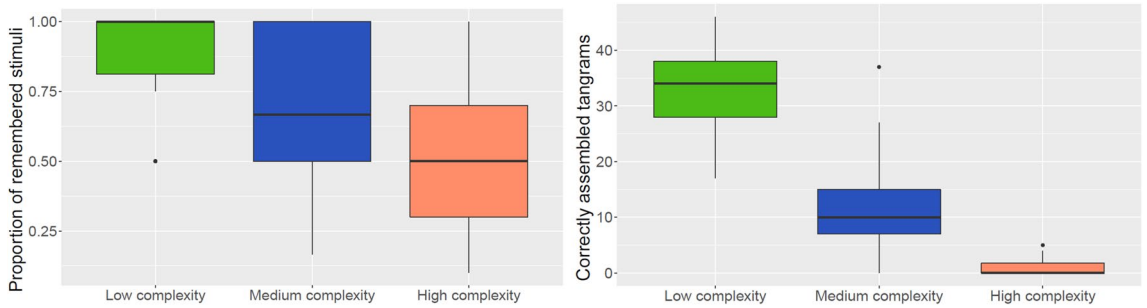

Fig. 4 The proportion of stimuli that participants remembered across the different conditions (information storage, figure on the left) and the number of tangram puzzles that participants correctly assembled (information processing, figure on the right)

Table 3 Correlation matrix showing relationships (Pearson correlation) between the self-reported cognitive load and the EDA measures and skin temperature, and between these physiological measures themselves

Measure	1	2	3	4	5	6	7
1. Cognitive load	–						
2. Mean skin temperature	− 0.15	–					
3. EDA: tonic skin conductance	0.05	− 0.06	–				
4. EDA: phasic skin conductance	0.11	0.04	0.81***	–			
5. EDA: skin conductance response rate	0.18	0.08	0.56***	0.81***	–		
6. EDA: skin conductance response duration	− 0.05	0.14	0.58***	0.74***	0.80***	–	
7. EDA: skin conductance magnitude	0.01	0.05	0.64***	0.82***	0.77***	0.93***	–

Significance codes: ***$p < .001$, **$p < 0.01$, *$p < 0.05$, '.' $p < 0.10$; $df = 136$

4.1.1 Evaluation of the physiological measures as indicators

The correlation matrix in Table 3 displays the relatedness between skin temperature, EDA measures and the self-reported cognitive load, together with the Pearson correlation coefficients.

The numbers in the first column indicate that none of the bivariate correlations between these measures and cognitive load is significant. However, as previously mentioned, care should be taken when interpreting these bivariate correlations, as they do not account for the repeated nature of the measures. Meanwhile, EDA measures seem to interrelate well.

The skin temperature initially correlated with the EDA measures, but that correlation disappeared upon removal of two outliers.

To have a first view on how EEG and EOG measures interrelate with each other and with participants' self-reported cognitive load, a second correlation matrix is depicted in Table 4. Although it is only a preliminary indication, one can see that the strongest correlations with cognitive load arise from the eye blink rate and the alpha peak frequency.

Multilevel analyses can alleviate the aforementioned flaws of bivariate correlations and are elaborated in the next paragraph. An analysis is made for EDA, EEG and EOG measures separately, and then for a combination of these measures.

4.2 Multilevel analyses

4.2.1 EDA measures and skin temperature as indicators of cognitive load

Each physiological measure monitored by the wrist-worn wearable is included in a separate multilevel analysis. These measures are the tonic and phasic component of skin conductance, the rate of skin conductance responses, the duration and magnitude of these skin conductance responses and the skin temperature. Results from these separate multilevel analyses show that none of these measures has a significant effect on the self-reported cognitive load (for each effect, $p > 0.05$). However, when analysing the five most distinct EDA measures (VIF < 10) together in a multilevel model, the skin conductance response duration ($p = 0.002$, $\beta = -0.002$) and the skin conductance response rate ($p < 0.001$, $\beta = 388$) are found to be significant (see Table 5). When combined, these five measures explain 11.5% of the variance in self-reported cognitive load. Note that the acceleration of the participants' left wrist has a highly significant effect ($p < 0.001$) on the self-reported cognitive load, but is not withheld in the analysis as it is a confounding factor that results from the design of the study: a more difficult condition automatically resulted in participants completing less puzzles, causing less movement of the wrist. These results provide some evidence for an association between the skin conductance response duration and response rate on the one hand and cognitive load on the other hand. The sizes of these effects, however, are small.

Table 4 Correlation matrix showing relationships between the self-reported cognitive load and the alpha power, alpha peak frequency and eye blink rate, and these physiological measures themselves

Measure	1	2	3	4	5
1. Cognitive load	–				
2. ERP: N200 voltage difference	− 0.05	–			
3. Alpha absolute power (Log, POz)	− 0.12	− 0.28**	–		
4. Alpha Peak frequency (Log, POz)	0.17*	− 0.24**	0.23**	–	
5. Eye blink rate	− 0.38***	0.03	0.05	− 0.03	–

Significance codes: ***$p < 0.001$, **$p < 0.01$, * $p < 0.05$, '.' $p < 0.10$; $df = 136$

Table 5 Results for the multilevel analyses of participants' self-reported cognitive load

	Null model		Model consisting of EDA measures and skin temperature		Model consisting of EEG measures		Model consisting of EOG measures		Model combining the best indicators across all types	
Fixed effects	β(SE)	*p*	β(SE)	*p*	β(SE)	*p*	β(SE)	*p*	β(SE)	*p*
Intercept	65 (2.1)	< .001***	63 (2.9)	< .001***	65(2.1)	< .001***	65 (1.9)	< .001***	65 (1.9)	< .001***
Mean skin temperature	–	–	− 4.0 (3.8)	.29	–	–	–	–	–	–
EDA: tonic skin conductance	–	–	0.5 (3.0)	.87	–	–	–	–	–	–
EDA: skin conductance response duration	–	–	− 0.002 (0.0006)	.002**	–	–	–	–	− 0.001 (0.0006)	.05
EDA: skin conductance response rate	–	–	388(104)	< .001***	–	–	–	–	239(105)	.02*
Alpha absolute power (Log, POz)	–	–	–	–	− 2.0(1.1)	.08	–	–	− 1.3(1.0)	.20
Alpha Peak frequency (Log, POz)	–	–	–	–	4.4 (2.3)	.06	–	–	4.5(2.0)	.03*
ERP: N200 voltage difference	–	–	–	–	− 0.2 (0.3)	.53	–	–	–	–
Eyeblink rate	–	–	–	–	-	-	− 0.56 (0.13)	< .001***	− 0.46 (0.13)	< .001***
Random effects										
Within subjects (residual) variance	444.7		393.5		429.5		375.6		343.4	
Between subjects variance	0		0		0		0		0	
Proportion of explained variance (R^2) [%]	Reference		11.5		3.4		15.5		22.8	
Model fit: AIC	916.4		907.8		915.8		900.2		894.9	

The different columns represent different models that include as physiological measures respectively EDA measures and skin temperature, EEG measures, EOG measures and a model consisting of a combination of the most significant physiological measures across all modalities

Significance codes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, '.' $p < 0.10$

4.2.2 EEG measures as indicators of cognitive load

Results from a multilevel analysis of EEG measures are depicted in Table 5. They indicate that the event-related N200 potential, i.e. the negative voltage difference assessed 200 ms after the initiation of deviating beep tones, does not relate to the self-reported cognitive load ($p > 0.05$). The logarithm of the alpha power has nearly a significant effect on the self-reported cognitive load ($p = 0.08$, $\beta = -2.0$). The logarithm of the alpha peak frequency has a positive and also nearly significant effect on the self-reported cognitive load ($p = 0.06$, $\beta = 4.4$). When combined, these two measures explain 3.4% of the variance in self-reported cognitive load. These nearly significant associations provide weak evidence for an increase in cognitive load to be associated with a lower alpha power and a higher alpha peak frequency.

4.2.3 EOG (eye blink rate) as an indicator of cognitive load

When analysing participants' eye blink rate, a clear significant negative association is found ($p < 0.001$, $\beta = -0.56$), in the sense that participants blinked their eyes less frequently when they reported a higher cognitive load (Table 5 shows the results). This measure allows to explain 15.5% of the variance in cognitive load. Of all investigated effect sizes, the strongest association is established between cognitive load and eye blink rate.

4.2.4 A model combining all data: EDA, skin temperature, EEG and EOG measures as indicators of cognitive load

Finally, a multilevel model is built consisting of the physiological measures that are most explanatory for cognitive load, across all "types": the skin conductance response duration and response rate, the logarithm of the alpha power, the logarithm of the alpha peak frequency and the eye blink rate.

The results from this multilevel analysis (last column of Table 5) show that with increasing cognitive load, participants' skin conductance response rate increases ($p = 0.02$), and the response rate durations decrease ($p = 0.05$). Alpha power is on average lower when cognitive load increases, but this is not significant. In addition, the frequency within the alpha power spectrum with the highest power increases with increasing cognitive load ($p = 0.03$). Finally, there is strong evidence for the rate of endogenous eye blinks to decrease with increasing cognitive load ($p < 0.001$). These five predictors can explain 22.8% of the variance in self-reported cognitive load. In sum, these results yield evidence for the cognitive load to be manifested by several physiological measures. The size of the different effects, however, is rather small: the majority of the variance in cognitive load can still not be explained through these measures.

5 Discussion

5.1 Implications for research

This study investigates whether and how well self-reported cognitive load can be measured through psychophysiological data. For that purpose, a controlled lab-setting inducing different levels of cognitive load is set up. The skin conductance response duration and response rate, the alpha power, the alpha peak frequency and the eye blink rate are identified as the best physiological markers for the cognitive load. However, they can only explain a limited proportion of the variance in cognitive load (22.8%). This limits the usability of EDA, EEG and EOG measures as measurement instruments for the cognitive load.

This study's results (Table 5) are partly in line with previous work (Table 2), in that some of the previous studies also observed a parietal alpha activity suppression (Ryu and Myung 2005; Antonenko et al. 2010) and a blink rate decrease (Ryu and Myung 2005) with an increasing cognitive load. However, some studies obtained insignificant or mixed results for these measures (Marquart et al. 2015; Haapalainen et al. 2010). In addition, none of the studies mentioned in Table 2 established EDA measures as significant, whereas this study's findings indicate that participants' skin conductance response rate increases and the response rate durations decrease with increasing cognitive load. Finally, this study did not cover pupil and eye-related measures, but it is noteworthy that previous work provides strong evidence for an association between these measures and cognitive load (Krejtz et al. 2018; Marquart et al. 2015; Rosch and Vogel-Walcutt 2013). Similarly, previous work also resulted in some evidence for an increase in cognitive load to be associated with an increase in heart rate variability (Solhjoo et al. 2019), a measure which this study could not cover.

Presumably, these differences can mainly be attributed to the limited statistical power of previous studies, especially because the effect sizes under study are probably inherently small. Next to that, differences in task design and the way in which cognitive load is inquired might influence how the physiological measures relate to the cognitive load scores.

This study also shows that a multimodal approach that includes multiple physiological markers is useful to increase the accuracy of the measurement. The more physiological markers that are included, the larger the proportion of variance that can be explained.

Next to that, the stringent methodological approach including the within-subjects design and the relatively large sample size have resulted in a relatively accurate parameter estimation, making the evidence about

associations between physiological measures and cognitive load much stronger.

This study also deliberately inquires the induced cognitive load. This entails a contribution from a methodological perspective, as it allows to account for the cognitive load being induced by a certain condition being person-dependent. When analysing repeated measures data of which outcome variables systematically vary from person to person, multilevel models are recommended, as they are especially suited to handle such inter-individual differences, for example when a participant systematically scores higher than others. In addition, we have shown that these models are convenient to evaluate the proportion of explained variance.

Another important theoretical insight that this study emphasizes (see Table 1) is that if one wants to measure a latent variable, it is very important that there is a rather direct link between the manifest variables (the physiological measures) and the latent variable (cognitive load). Without a theoretical underpinning of the relation between a manifest variable and the latent variable, it is likely that no or a very weak association will be found.

We also recommend to include a cognitive overload condition as it is interesting from a methodological point of view, making associations between cognitive load and physiological data more visible. Next to that, cognitive overload is interesting to study as it negatively impacts performance and well-being (frustration, stress and burn-out) and is thus relevant to white- and blue-collar contexts (Young et al. 2014).

The existence of significant physiological indicators enables researchers to conduct cross-sectional studies on a group level to compare the effect of different conditions on cognitive load (i.e. different tasks, instructional designs, boundary conditions, etc.). Provided that lab-settings are sufficiently controlled and that the tested samples are sufficiently large, it will be likely to discover differences between conditions in terms of participants' cognitive load, if these are large enough when measuring these physiological signals.

However, as also concluded by Cranford et al. (2014) and by Fisher et al. (2018), significant findings across individuals do not automatically imply that accurate (real-time) measurements on an individual subject's level are possible. Despite the observed significant physiological markers, the majority of the variance (77.2%) in cognitive load cannot be explained. This implies that cognitive load cannot be measured accurately on a single subject by means of the physiological measures used in this study. It is important for researchers and practitioners to realize that shortcoming in order not to mistakenly overestimate the potential of measuring the cognitive load on a single subject, and to have a realistic view on possible applications in practice.

The findings from the performance measures seem straightforward and indeed confirm the theory that the human working memory and spatial ability are limited: the more information processing and storage is required, the lower the proportion of stimuli that can actually be remembered and the puzzles that can be completed.

5.2 Implications for practice

Developments of wearable sensors, increasing computational power and evolutions in information technology and in cognitive psychology may potentially lead to wearable devices that measure cognitive load. Applied to the example of assembly workers, one could think of identifying operators who frequently suffer from high cognitive load levels and may need more support or training, would be better assigned to other tasks, or need professional help in view of burn-out prevention.

However, the results of our multimodal approach and of much-related work (see Table 2) show that such accurate measurements on a particular single operator are not yet possible.

Next to measuring individual operators, one could also monitor and compare the cognitive load between groups of operators, for instance in view of evaluating and comparing assembly stations or new production methods. Our results indicate that such comparisons based on a group design should be possible. However, in an assembly context, it seems unlikely that there will be many cases in which the costs (measurement equipment, time and corresponding production loss) of this application will outweigh its benefits.

Note that this study's experimental task does not solely apply to assembly work. The study may be understood in a broader context of applications that consist of (visuo-spatial) information processing and storage.

Next to the accuracy of measuring, privacy is another hurdle when measuring the physiology of employees. As this is not the focus of the current study, this concern is not further elaborated here.

5.3 Limitations

This study is prone to several limitations. These limitations also represent the underlying reasons for the very limited obtained proportion of variance in cognitive load.

First, although this study considers multiple physiological measures, these still represent a selection. To be more specific,

this study does not consider eye-tracking nor pupillometry, although several studies (e.g., Krejtz et al. 2018) have shown the relationship between pupil dilatation and microsaccades magnitude on the one hand and cognitive load on the other hand. In a similar way, the heart rate and heart rate variability (see Ryu and Myung 2005) could not be analysed either. Including these and other relevant measures in the model may eventually further increase the proportion of explained variance.

A second limitation is that subjects are only measured during a relatively short timeframe. As subjects are not followed over a longer time span, it is not possible to include person-specific parameters to enhance the model fit.

A third limitation is that we used self-reports as the gold standard, although also this measure does not perfectly reflect the cognitive load. The validity of self-reports can be hampered by incorrectly interpreting the question and by the difficulty in retrospectively assessing one's own cognitive load. Note that Matthews et al. (2019) even claim that self-reports and psychophysiological measures are divergent, and conclude that "various available workload measures assess not one but several distinct constructs" (p. 20). They attribute the reasons for this divergence to several causes, such as deficiencies in subjective measurement scales, absence of the latent construct of interest, deficiencies in objective measures themselves or workload being non-unitary. Another drawback to the validity of self-reporting is that we inquired cognitive load by means of a unidimensional approach (similarly as Paas 1992), and not via multiple items.

5.4 Future work

Future work could include extending the multimodal approach by including heart rate measures and pupillometry. Next to that, it could be interesting to investigate how and how well cognitive load could be measured in less controlled contexts, such as in factory environments. Note that combining EEG, EOG, ECG and pupillometry poses several new challenges when moving these techniques from a lab environment to less controlled environments in the "real world". An obstacle particularly related to pupillometry is that this technique is not adequate yet for in-the-field usage, mainly due to variance in luminance coming from for instance the factory environment (assembly components, work table, etc.) (Van Acker et al. 2020). Finally, the person-specific nature of psychophysiological encourages new research lines in which longitudinal designs are set up, in which subjects are measured on multiple occasions during longer time spans to examine whether personalised models can enhance the proportion of variance in the cognitive load that can be explained (in a similar line of thought as e.g. Haapalainen et al. 2010).

6 Conclusion

The first research aim of this study was to investigate how well cognitive load can be measured through physiological data. The results highlight that finding significant markers across individuals does not automatically imply that accurate measurements on an individual level are possible. The skin conductance response duration and response rate, the alpha power, the alpha peak frequency and the eye blink rate are identified as significant markers for cognitive load, but together, they can only explain 22.8% of its variance.

The second research aim was to evaluate the corresponding implications both for research and for practice. Results show that the multimodal approach addressed in this study does not enable to measure cognitive load in an accurate way.

A first way to try to improve the measurement in future work is by extending the multimodal approach, by including eye-tracking, pupillometry and heart rate variability. A second way may be to collect more longitudinal measurements and consider personalised models that allow the way in which cognitive load manifests itself in a physiological variable (i.e., the regression coefficients) to differ from person to person.

Improving the measurement model and re-evaluating its accuracy is necessary before even starting to consider applications in practice.

Acknowledgements This work was executed within a one-year research project funded by imec, executed by different research teams affiliated to imec. In this project, mict's expertise related to physiology (EEG, EOG and EDA) and Itec's expertise related to instructional design and statistical modelling is brought together.

References

Antonenko P, Paas F, Grabner R, van Gog (2010) Using electroencephalography to measure cognitive load, Educ Psychol Rev 22(4), 425-438. Doi: 10.1007/s10648-010-9130-y

Ayaz H, Shewokis PA, Bunce S, Izzetoglu K, Willems B, Onaral B (2012) NeuroImage Optical brain monitoring for operator training and mental workload assessment. NeuroImage 59(1):36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023

Belletier C, Charkhabi M, Silva GPA, Ametepe K, Lutz M, Izaute M (2019) Wearable cognitive assistants in a factory setting: a critical review of a promising way of enhancing cognitive performance and well-being. Cognition Technol Work. https://doi.org/10.1007/s10111-019-00610-2 (**Advanced online publication**)

Bianchi R, Schonfeld IS, Laurent E (2015) Burnout—depression overlap: a review. Clin Psychol Rev 36:28–41. https://doi.org/10.1016/j.cpr.2015.01.004

Brouwer A, Zander TO, Van Erp JBF, Korteling JE, Bronkhorst AW (2015) Using neurophysiological signals that reflect cognitive or affective state : six recommendations to avoid common pitfalls. Front Neurosci 9:136. https://doi.org/10.3389/fnins.2015.00136

Chen F, Zhou J, Wang Y, Yu K, Arshad S, Khawaji A, Conway D (2016) Robust multimodal cognitive load measurement. Hum Comput Interaction Ser Cham. https://doi.org/10.1007/978-3-319-31700-7 (**Springer International Publishing**)

Cranford KN, Tiettmeyer JM, Chuprinko BC, Jordan S, Grove NP (2014) Measuring load on working memory: the use of heart rate as a means of measuring chemistry students ' cognitive load. J Chem Educ 91:641–647. https://doi.org/10.1021/ed400576n

El Maraghy W, El Maraghy H, Tomiyama T, Monostori L (2012) Complexity in engineering design and manufacturing. CIRP Ann Manuf Technol 61(2):793–814. https://doi.org/10.1016/j.cirp.2012.05.001

European Commission, Executive Agency for Small and Medium-sized Enterprises (EASME), PwC (2019) curriculum guidelines for key enabling technologies (kets) and advanced manufacturing tcBorders="YES"technologies (AMT): Belgium. https://op.europa.eu/en/publication-detail/-/publication/4dcaeee3-29c2-11e9-8d04-01aa75ed71a1/language-en/format-PDF/source-87225354x

Fisher AJ, Medaglia JD, Jeronimus BF (2018) Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci USA 115(27):6106–6115. https://doi.org/10.1073/pnas.1711978115

Fryer L, Dinsmore D (2020) The promise and pitfalls of self-report. Frontline Learn Res 8(3):1–9. https://doi.org/10.14786/flr.v8i3.623

Haapalainen E, Kim S, Forlizzi JF, Dey AK (2010) Psycho-Physiological Measures for Assessing Cognitive Load. In: Psychophysiological measures for assessing cognitive load, 12th ACM international conference on ubiquitous computing (pp 301–310). Copenhagen. https://www.cs.cmu.edu/~sjunikim/publications/UBICOMP2010_Cognitive_Load.pdf

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Advances in psychology, 52, Human mental workload, Oxford, England: North-Holland, pp 139–183 doi: 10.1016/S0166-4115(08)62386-9

Herborn K, Graves J, Jerem P, Evans N, Nager R, Mccafferty D, Mckeegan D (2015) Skin temperature reveals the intensity of acute stress. Physiol Behav 152:225–230. https://doi.org/10.1016/j.physbeh.2015.09.032

Iskander M (2018) Burnout, cognitive overload, and metacognition in medicine. Med Sci Educ 29(1):325–328. https://doi.org/10.1007/s40670-018-00654-5

Jerčić P, Sennersten C, Lindley C (2018) Modeling cognitive load and physiological arousal through pupil diameter and heart rate. Multimedia Tools Appl. https://doi.org/10.1007/s11042-018-6518-z

Johannsen G (1979) Workload and workload measurement. In: Moray N (ed) Mental workload: Its theory and measurement. Springer, Boston, pp 3–11

Kramer AF (1991) Physiological metrics of mental workload: a review of recent progress. In: Damos DL (ed) Multiple-task performance. Taylor and Francis, London, pp 279–328

Krejtz K, Duchowski AT, Niedzielska A, Biele C, Krejtz I (2018) Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. PLoS ONE 13(9):e0203629. https://doi.org/10.1371/journal.pone.0203629

Kyllonen PC, Christal RE (1990) Reasoning ability is (little more than) working-memory capacity?! Intelligence 4:389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Lasi H, Fettke P, Kemper H, Feld T, Hoffmann M (2014) Industry 4.0. Busin Inform Syst Eng 6(4):239–242

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 Learn Res 7(2):57–74. https://doi.org/10.14786/flr.v7i2.403

Ledger H (2013) The effect cognitive load has on eye blinking. Plymouth Stud Scientist 6(1), 206–223. https://bcur.org/journals/index.php/TPSS/article/view/373

Leppink J, Paas F, Van der Vleuten CPM, Van Gog T, Van Merriënboer JJG (2013) Development of an instrument for measuring different types of cognitive load. Behav Res Methods 45:1085–1092. https://doi.org/10.3758/s13428-013-0334-1

Liu Y, Ayaz H, Shewokis PA (2017) Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy. Brain-Comput Interf 4(3):175–185. https://doi.org/10.1080/2326263X.2017.1304020

Luck SJ (2012) Event-related potentials. In: H Cooper, PM Camic, DLLong, AT Panter, D Rindskopf, KJ Sher (eds) APA handbooks in psychology® APA handbook of research methods in psychology, Vol 1 Foundations, planning, measures, and psychometrics (pp 523–546). American Psychological Association. https://doi.org/10.1037/13619-028

Matthews G, De Winter J, Hancock PA (2019) What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. Theoretical Issues Ergonomics Sci. https://doi.org/10.1080/1463922X.2018.1547459

Marquart G, Cabrall C, de Winter J (2015) Review of eye-related measures of drivers' mental workload. In: Proceedings of the 6th international conference on applied human factors and ergonomics (AHFE 2015) and the affiliated conferences, AHFE 2015 (3, 2854–2861). https://doi.org/10.1016/j.promfg.2015.07.783

Morton J, Vanneste P, Larmuseau C, Van Acker BB, Raes A, Bombeke K, Cornillie F, Saldien J, 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 2019). Rome, Italy. https://arrow.tudublin.ie/hwork19/1/

Noroozi O, Alikhani I, Järvelä S, Kirschner P, Juuso I, Seppänen T (2019) Multimodal data to design visual learning analytics for understanding regulation of learning. Comput Hum Behav 100:298–304

Paas F (1992) Training strategies for attaining transfer of problem solving skills in statistics: a cognitive load approach. J Educ Psychol 84:429–434

Pekrun R (2020) Self-report is indispensable to assess students' learning. Frontline Learn Res 8(3):185–193. https://doi.org/10.14786/flr.v8i3.637

Reid GB, Nygren TE (1988) The subjective workload assessment technique: a scaling procedure for measuring mental workload. In: Hancock PA, Meshkati N (eds) Advances in psychology, 52. Human mental workload (pp 185–218). North-Holland. https://doi.org/10.1016/S0166-4115(08)62387-0

Richardson M, Jones G, Torrance M, Baguley T (2006) Identifying the task variables that predict object assembly difficulty. Hum Factors 48(3):511–525

Rosch JL, Vogel-Walcutt JJ (2013) A review of eye-tracking applications as tools for training. Cogn Technol Work 15(3):313–327. https://doi.org/10.1007/s10111-012-0234-7

Ryu K, Myung R (2005) Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int J Ind Ergon 35:991–1009. https://doi.org/10.1016/j.ergon.2005.04.005

Schnotz W, Kürschner C (2007) A reconsideration of cognitive load theory. Educ Psychol Rev 19(4):469–508. https://doi.org/10.1007/s10648-007-9053-4

Schonfeld IS, Bianchi R (2016) Burnout and depression: two entities or one ? J Clin Psychol 72(1):22–37. https://doi.org/10.1002/jclp.22229

Setz C, Arnrich B, Schumm J, La Marca R, Troster G, Ehlert U (2010) Discriminating stress from cognitive load using a wearable EDA device. IEEE Trans Inf Technol Biomed 14(2):410–417. https://doi.org/10.1109/TITB.2009.2036164

Solhjoo S, Haigney MC, McBee E, Van Merrienboer JJG, Schuwirth L, Artino ARAJ, Battista A, Ratcliffe TA, Lee HD, Durning SJ (2019) Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. Sci Rep 9(1):1–9. https://doi.org/10.1038/s41598-019-50280-3

Sweller J (1988) Cognitive load during problem solving: effects on learning. Cognitive Sci 12(1):257–285. https://doi.org/10.1016/0364-0213(88)90023-7

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

Sweller J, Van Merrienboer JJG, Paas FGWC (1998) Cognitive architecture and instructional design. Educ Psychol Rev 10(3):251–296. https://doi.org/10.1023/A:1022193728205

Sweller J (2010) Cognitive load theory: Recent theoretical advances. In: Plass JL, Moreno R, Brünken R (eds) Cognitive load theory, Cambridge University Press, Cambridge, pp 29–47

Van Acker BB, Parmentier DD, Vlerick P, Saldien J (2018) Understanding mental workload: from a clarifying concept analysis toward an implementable framework. Cogn Technol Work 20(3):351–365. https://doi.org/10.1007/s10111-018-0481-3

Van Acker BB, Bombeke K, Durnez W, Parmentier DD, Mateus JC, Biondi A, Saldien J, Vlerick P (2020) Mobile pupillometry in manual assembly: a pilot study exploring the wearability and external validity of a renowned mental workload lab measure. Int J Indu Ergonomics, 75. https://doi.org/10.1016/J.ERGON.2019.102891

Van der Wel P (2018) Van Steenbergen H (2018) Pupil dilation as an index of effort in cognitive control tasks: A review. Psychonomic Bulletin Review 25:2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Young MS, Brookhuis KA, Wickens CD, Hancock PA (2014) State of science : mental workload in ergonomics. *Ergonomics, 58*(1). https://doi.org/10.1080/00140139.2014.956151

Wan X, Sanders NR (2017) The negative impact of product variety: forecast bias, inventory levels, and the role of vertical integration. Int J Prod Econ 186:123–131. https://doi.org/10.1016/j.ijpe.2017.02.002

Zagermann J, Pfeil U, Reiterer H (2018) Studying eye movements as a basis for measuring cognitive load. In: *Conference on human factors in computing systems (pp. 1–6)*. Montréal, Canada. Doi: 10.1145/3170427.3188628

Zheng Y, Ding X, Poon CCY, Lo BPL, Zhang H, Zhou X, Yang G, Zhao N, Zhang Y (2014) unobtrusive sensing and wearable devices for health informatics. IEEE Trans Biomed Eng 61(5):1538–1554. https://doi.org/10.1109/TBME.2014.2309951

Zimmerman BJ (2008) Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. Am Educ Res J 45(1):166–183