Exploring the Validity of Physiological Measures for Cognitive Load Assessment in Virtual Reality

Laurent Lacroix, Olivier Augereau, Nathalie Le Bigot

Exploring the Validity of Physiological Measures for Cognitive Load Assessment in Virtual Reality

Laurent Lacroix
Université de Bretagne Occidentale
Brest, France
laurent.lacroix@etudiant.univ-brest.fr

Olivier Augereau
Lab-STICC CNRS UMR 6285
École Nationale d'Ingénieurs de Brest
Brest, France
augereau@enib.fr

Nathalie Le Bigot
Lab-STICC CNRS UMR 6285
Université de Bretagne Occidentale
Brest, France
nathalie.lebigot@univ-brest.fr

ABSTRACT

Cognitive load triggers the researchers' interest in various fields. In the context of education and training, maintaining an optimal cognitive load is crucial to keep learners engaged, ensuring that the content aligns with their skill levels. However, assessing cognitive load is challenging. In the literature, several methods have been proposed mainly through questionnaires, performance metrics and physiological sensors.

In this paper, we propose an experiment in virtual reality where four different tasks has been designed to stimulate different levels of cognitive load. We compare three different ways of measuring the cognitive load to estimate the validity of physiological measures. The findings suggest that, to some extent, the physiological measures are well-suited for assessing cognitive load in the context of this study. This research contributes valuable insights to the ongoing exploration of effective cognitive load measurement methodologies.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Virtual reality**; **Empirical studies in HCI**; • **Applied computing** → **Psychology**;

KEYWORDS

Cognitive Load, Virtual Reality, Physiological Measures

ACM Reference Format:

Laurent Lacroix, Olivier Augereau, and Nathalie Le Bigot. 2018. Exploring the Validity of Physiological Measures for Cognitive Load Assessment in Virtual Reality. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Cognitive load has been a focal point of extensive research across various disciplines such as psychology, ergonomics, and computer science. The foundation of cognitive load theory is rooted in the limitations of working memory. When an individual engages in a task that requires mental effort, utilizing their working memory—such as solving a problem or following instructions—it induce a load on the working memory system [9].

The concept of cognitive load suggests that each task incurs a specific cost, corresponding to the cognitive load it imposes on the individual. As individuals navigate through various tasks, especially in multitasking scenarios, they must allocate a portion of their cognitive resources to each concurrent task. This allocation increases the overall cognitive load experienced by the individual and can potentially lead to cognitive overload. In situations of cognitive overload, observable outcomes may include notable slowdowns in task execution and, in some cases, the inability to successfully complete one or more tasks [19].

Understanding cognitive load is crucial for designing effective interfaces, educational materials, and work environments. By considering the cognitive load associated with different tasks, researchers and practitioners can develop strategies to optimize cognitive resources, enhance task performance, minimize the risk of cognitive overload and propose interactive virtual environment.

1.1 Cognitive load measurement

Cognitive load measures are commonly categorized into two main types: subjective and objective measures.

Subjective measures revolve around the user's own perception of task difficulty and the mental effort invested. These measures are typically acquired through questionnaires such as the NASA Task Load Index (NASA-TLX) developed by Hart and Staveland [6], the Subjective Workload Assessment Technique (SWAT) introduced by Reid and Nygren [14] and the Instantaneous Self-Assessment (ISA) scale [20] proposed by Tattersall and Foord. These scales help capture the user's subjective experience of cognitive load and task difficulty. However, self-reported measures have been criticized to have potential biases [8] and cannot be used in real time to adapt a virtual environment to the user's need.

On the other hand, objective measures can be obtained through performances [5] or physiological signals [7]. Performance measures gauge how effectively a user accomplishes a given task. These can include traditional metrics such as reaction time to a stimuli [3] or the number of errors performed while doing a task [2]. A secondary task is also often employed in dual-task paradigms [16] to stimulate cognitive load and to gather additional objective measures. However performance measures are quite specific for a task, and not all tasks are designed to collect such measures.

Physiological measures offer insights into cognitive load by monitoring the user's physiological responses. Research has demonstrated that certain physiological indicators are reliable indicators

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

of cognitive load. For instance, eye movements can reflect the reader understanding [10], electrodermal activity can signal stress during task execution [7], heart rate and heart rate variability may also provide valuable information about cognitive workload [18]. Physiological-based cognitive load assessment has the advantage to be done in real time and to be usable in various situations.

The integration of both subjective and objective measures in cognitive load assessment provides a comprehensive understanding of the user's cognitive experience. Combining insights from users' self-reported perceptions with physiological and performance data will provide a more nuanced and holistic view of cognitive load.

1.2 Physiological-based measurement

Physiological responses are indicative of cognitive load [4]. Several kind of signals have been investigated for measuring cognitive load, especially: photoplethysmography (PPG) [11, 17], eletrocardiography (ECG) [15], electrodemal activity (EDA) [11], electroencephalography (EEG) [15, 21], functional Near-Infrared Spectroscopy (fNIRS) [13], respiration [12], eye-tracking [11, 17], etc.

The most recent methods are usually based on machine learning algorithms trained on users physiological data while they are engaged in tasks with varying levels of difficulty. For our experiment we selected the method proposed by Siegl et al. [17], as their cognitive load algorithm is accessible through an sdk[1].

In the presented experiment, our goal is to validate the accuracy and applicability of the cognitive workload measurements obtained through this physiological model across various conditions. To achieve this, we used an open-source platform [1] featuring four virtual environments designed to elicit distinct levels of cognitive load. Our hypothesis posits that the accuracy of the cognitive load measurement should vary based on the task difficulty and exhibit correlations with both subjective and objective measures.

2 METHOD AND MATERIALS

2.1 Participants

In this study, 19 voluntary participants, aged between 21 and 41 years old (mean = 23.9), were recruited. The sample comprised 10 females and 9 males, all with normal or corrected-to-normal vision. Participants self-assessed their proficiency in new technologies on a scale from 3/10 to 10/10 (mean = 6.79). Throughout the experiment, participants were required to respond to audio stimuli. However, the data from one participant were excluded from the analysis due to that participant missing more than 50% of the audio reaction test.

2.2 Material

We utilized the "HP Reverb G2 Omnicept Edition" headset to display the environment in VR and to measure users' cognitive load. The machine learning system developed by Siegel et al. [17] predicts cognitive load based on heart rate, eye position and openness, and pupil diameter. To induce cognitive load, we employed an open-source virtual environment [1]. Participants found themselves in a minimalist virtual room with a prominently displayed numeric keypad.

[1]https://developers.hp.com/omnicept/downloads

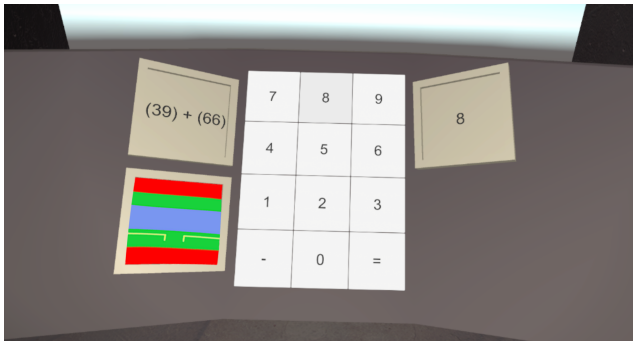

Figure 1: The virtual environment used for the experiment. The numeric keypad is used to answer the mathematical operation displayed in the top left panel. The top right panel displays the current response of the user before validation. The bottom left panels appeared only in the second scene. In this panel a white line is moving in front of a gauge.

This virtual environment comprises two distinct scenes. In the initial scene, two panels were presented on either side of the keypad. The top left panel showcased mathematical operations for participants to solve, while the top right panel displayed the current answer. In the second scene, the arrangement remained the same, with an additional panel appearing beneath the mathematical operations. This extra panel featured a gauge with a moving line, challenging participants to keep the line within a predefined threshold (Figure 1). The NASA-TLX questionnaire was integrated into the virtual environment between different tasks to assess the participant's cognitive load (Figure 2).

2.3 Procedure

Participants experienced four distinct conditions, with the initial two (C1 and C2) occurring in the first scene, and the subsequent two (C3 and C4) taking place in the second scene. At the beginning of each scene, participants underwent a one-minute phase aimed at familiarizing themselves with the virtual environment.

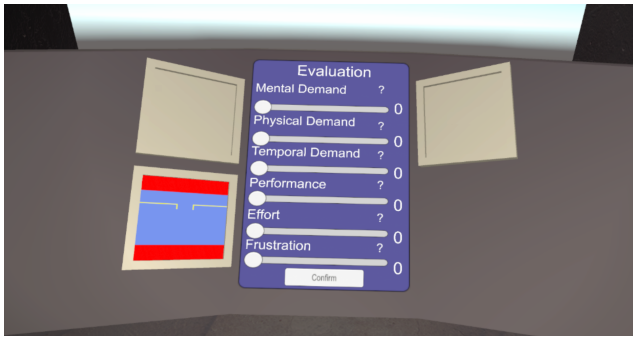

Figure 2: After each task, the NASA-TLX is displayed in the virtual environment. The users can click on the interrogation marks to display the questions related with each item.

The scenes followed a uniform structure: one condition lasting 3 minutes, a 2-minute break during which participants responded to the NASA-TLX questionnaire, another 3-minute condition, and a subsequent NASA-TLX questionnaire regarding the second condition. After the first scene, participants were instructed to remove the headset, allowing for a brief intermission before commencing the second scene. The sequence of scenes was balanced among participants.

2.4 Conditions and measures

Throughout each condition, participants were tasked with promptly responding to audio stimuli by pressing a button on one controller while simultaneously interacting with the virtual environment using the other controller. We will now detail each of the four conditions.

(C1) In the first condition of the first scene, participants were tasked with responding to "simple" calculations, involving the summation of two-digit numbers. To answer the calculations, they had to use the right controller to point at different digits.

(C2) In the second condition of the first scene, participants faced more challenging calculations, requiring the summation of three-digit numbers randomly selected from the range of 100 to 200.

(C3) In the first condition of the second scene, participants were instructed to control a moving line within a specified interval. The line moved vertically, and pressing a button on the controller changed its direction to prevent it from reaching a predefined zone (Figure 1). No calculations were involved in this condition.

(C4) In the second condition of the second scene, participants encountered a dual-task scenario. They were required to both maintain the line within the interval and respond to simple calculations simultaneously.

Finaly, the cognitive load was assessed through four measures: (1) the headset "cognitive load" score based on physiological sensors varying between 1 (high cognitive load) and 0 (low cognitive load); (2) the average response time to the audio stimuli; (3) the number of correctly solved calculations; and (4) the NASA-TLX average scores.

3 RESULTS

We conducted repeated measures ANOVA on the cognitive load recorded by the headset in the four conditions. The ANOVA showed that the effect of conditions on cognitive load is significant: $F(3, 51) = 39.012; p < .001$ (Figure 3). Post hoc tests showed that cognitive load is not significantly different between C1 and C2 ($p_{bonf} = .200$) but it is significantly lower in C3 as compared to C2 ($p_{bonf} < .001$) and significantly higher in C4 as compared to C1($p_{bonf} = .037$).

The same repeated measures ANOVA was conducted on the NASA-TLX scores. The ANOVA showed that the effects of conditions on estimated cognitive load is also significant ($F(3, 51) = 53.752; p < 0.001$) (Figure 4). Post hoc tests revealed that only C2 and C4 are not significantly different with $\alpha = .05$ ($p_{bonf} = .066$).

Concerning audio response times, the data which lies beyond ±2.5 times the standard deviation from the mean, was considered as outlier and excluded. Then we conducted repeated measures ANOVA on the average response time. It showed a main effect of conditions on response time ($F(1.369, 23.265) = 12.794; p < .001$)

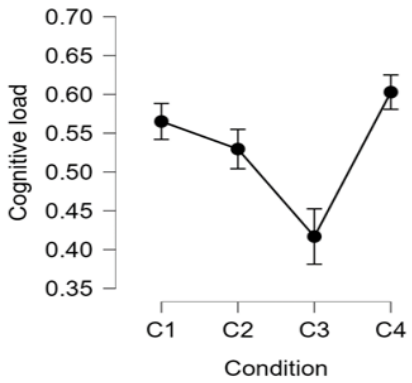

Figure 3: Mean cognitive load measures based on physiological measures provided by the headset, for each condition. Error bars represent the 95% confidence intervals.

(Figure 5). Here, post hoc tests showed that response time is only significantly higher in C4 ($p_{bonf} < .01$, for each comparison). We applied the Greenhouse-Geisser correction on the data sample, because the sphericity condition on this sample was not respected.

The last repeated measures ANOVA on successfully realized calculations showed that the effect of conditions on realized calculations is significant ($F(1.509, 25.657) = 15.181; p < .001$). Post hoc tests revealed that all conditions are significantly different when using Bonferroni correction with $\alpha = .05$. The sphericity condition was not respected so we also applied the Greenhouse-Geisser on the sample.

In a second step, we calculated correlation between different measures in order to verify which measures were correlated with the cognitive load measured from physiological data. Results showed that the headset's measure is positively correlated with audio response time ($r = .318; p = .007$) and with NASA-TLX mean scores ($r = .538; p < .001$). The number of realized calculations is not correlated with cognitive load ($r = -.156; p = .261$).

We tested correlations between our own measures too. NASA-TLX mean scores are correlated positively with audio response time ($r = .412; p < .001$), and correlated negatively with realized calculation ($r = (-0.512); p < .001$). Response time and number of calculations realized are also correlated negatively ($r = (-0.376); p = .005$).

4 DISCUSSION

The ANOVA on the cognitive load provided by the headset revealed significant differences between conditions. It means that the headset is currently detecting cognitive load variations depending on the task. Moreover, the results follow our expectation for the conditions C1, C3 and C4. C3 is very simple and the cognitive load associated with it is the lowest. C4 was designed to be the hardest with a dual task, and the measured cognitive load is the highest. However, we observed that for C1 and C2, the cognitive load is relatively similar. We designed the experiment by expecting that C1 would be easier than C2 but it might not be the case. If we check the other measures, it is also not clear. The NASA-TLX average score is lower for C1 than C2 but the average response time is lower for C2 than C1. The

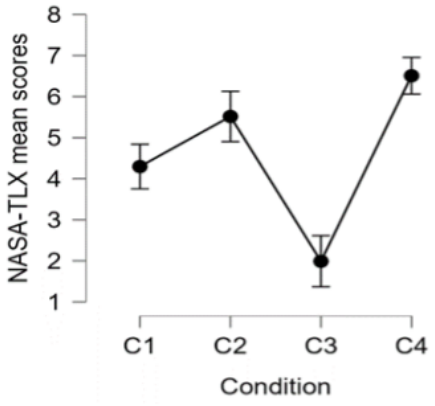

Figure 4: Mean NASA-TLX scores for each condition. Error bars represent 95% confidence intervals.

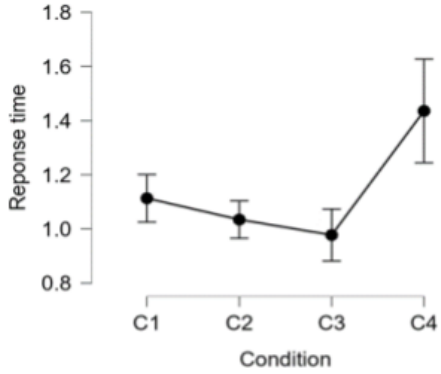

Figure 5: Mean response time for each condition. Error bars represent 95% confidence intervals.

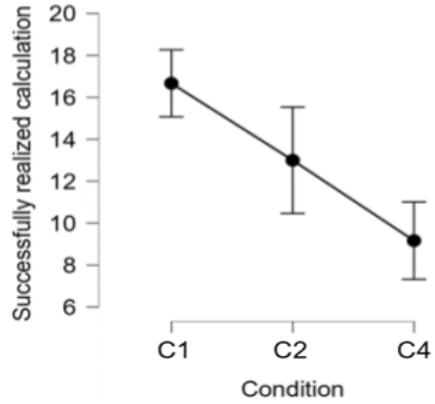

Figure 6: Mean number of calculations successfully realized in each condition. No calculations are done in C3. Error bars shows 95% confidence intervals.

number of successfully realized calculation is lower but it might be simply due to the fact that there is one more digit to consider and to input, which takes more time. But it does not necessarily increase the participant's cognitive load. Some participants even related that, in C2 the calculation with three digits was easier for them because the numbers were aligned on top of each other (and not in a line as in C1). The conclusion is that the designs of C1 and C2 was not different enough to capture significant different measures.

The ANOVA on Nasa-TLX mean scores revealed that subjective evaluation of cognitive load is following our expectations. Participants felt more difficulties in C4, which was the hardest. In the same way, mean scores are the lowest for C3. For the two first conditions, we can see that mean scores are higher for C2 than for C1.

For the audio reaction time, the ANOVA showed that only C4 is significantly different to the other conditions. The dual task scenario clearly induces a cognitive overload which is consistent with the design of the virtual environment. In a similar way to the headset measures, C3 seems to be the easiest with the lowest response time, C2 a little bit harder and C1 a little bit harder than C2. However these differences are not significant.

The last ANOVA on successfully realized calculation showed that in C4, the participants realized much less calculations than in C1. In both conditions the participants have to calculate the sum of double digit numbers, but in C4 an extra task was added which reduce the performances of the participants. For C3, triple digit calculation might not be much harder than the double one, but is simply more time consuming.

To conclude, all ANOVAs showed that C4 is the hardest, C3 the easiest. The unclear point concerns C1 and C2, which is probably the result of an non significant task complexity difference. Concerning correlations, nearly all the measures align with our expectations. As cognitive load, as measured by the headset, increases, both reaction time and NASA-TLX scores exhibit a corresponding increase. Specifically, as tasks become more challenging and demand greater cognitive resources, participants face increased difficulty during execution, resulting in decreased performance. The correlation with the number of completed calculations did not reach statistical significance, but there is an observable trend in that direction. It may be more meaningful in future studies to focus on measuring error rates. In the present experiment, the same calculation was displayed until a correct answer was provided; this make the assessment of error rates less applicable.

5 CONCLUSION & FUTURE WORK

In conclusion, the physiological measures based on the headset appear to be a relatively accurate indicator of cognitive load, given the consistency and correlation observed with both subjective and performance measures. The tasks, particularly C1, C3, and C4, were effectively designed to induce variations in cognitive load. However, results were less conclusive for C2, raising questions about the design of C2.

For future investigations, it is crucial to introduce more substantial differences in difficulty between each condition and mitigate presentation bias. It is also important to test these measures in different scenarios before firmly asserting the physiological data

usability across various tasks. Nonetheless, this study serves as a promising initial step towards that goal.

ACKNOWLEDGMENTS

This work was supported by the French government funding managed by the National Research Agency under the Investments for the Future program (PIA) grant ANR-21-ESRE-0030 (CONTINUUM).

REFERENCES

[1] Olivier Augereau, Gabriel Brocheton, and Pedro Paulo Do Prado Neto. 2022. An Open Platform for Research about Cognitive Load in Virtual Reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 54–55.

[2] Paul Ayres. 2006. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and instruction* 16, 5 (2006), 389–400.

[3] Pierre Barrouillet, Sophie Bernardin, Sophie Portrat, Evie Vergauwe, and Valérie Camos. 2007. Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 3 (2007), 570.

[4] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psychophysiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 301–310.

[5] Faizal A Haji, David Rojas, Ruth Childs, Sandrine de Ribaupierre, and Adam Dubrowski. 2015. Measuring cognitive load: performance, mental effort and simulation task complexity. *Medical education* 49, 8 (2015), 815–827.

[6] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[7] Curtis S Ikehara and Martha E Crosby. 2005. Assessing cognitive load with physiological sensors. In *Proceedings of the 38th annual hawaii international conference on system sciences*. IEEE, 295a–295a.

[8] Samuel C Karpen. 2018. The social psychology of biased self-assessment. *American Journal of Pharmaceutical Education* 82, 5 (2018).

[9] Paul A Kirschner, Femke Kirschner, and Fred Paas. 2009. Cognitive load theory. In *Psychology of classroom learning: An encyclopedia*. Macmillan Reference, 205–209.

[10] Charles Lima Sanches, Koichi Kise, and Olivier Augereau. 2017. Japanese reading objective understanding estimation by eye gaze analysis. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 121–124.

[11] Tiffany Luong, Nicolas Martin, Anais Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lécuyer. 2020. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 425–437.

[12] Lambertus JM Mulder. 1992. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology* 34, 2-3 (1992), 205–236.

[13] Felix Putze, Christian Herff, Christoph Tremmel, Tanja Schultz, and Dean J Krusienski. 2019. Decoding mental workload in virtual environments: a fNIRS study using an immersive n-back task. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3103–3106.

[14] Gary B Reid and Thomas E Nygren. 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*. Vol. 52. Elsevier, 185–218.

[15] Kilseop Ryu and Rohae Myung. 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics* 35, 11 (2005), 991–1009.

[16] Cornelia Schoor, Maria Bannert, and Roland Brünken. 2012. Role of dual task design when measuring cognitive load during multimedia learning. *Educational Technology Research and Development* 60 (2012), 753–768.

[17] EH Siegel, J Wei, A Gomes, M Oliviera, P Sundaramoorthy, K Smathers, M Vankipuram, S Ghosh, H Horii, J Bailenson, et al. 2021. *HP Omnicept cognitive load database (HPO-CLD)–developing a multimodal inference engine for detecting real-time mental workload in VR.* Technical Report. Technical Report.

[18] Soroosh Solhjoo, Mark C Haigney, Elexis McBee, Jeroen JG van Merrienboer, Lambert Schuwirth, Anthony R Artino Jr, Alexis Battista, Temple A Ratcliffe, Howard D Lee, and Steven J Durning. 2019. Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific reports* 9, 1 (2019), 14668.

[19] John Sweller. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review* 22 (2010), 123–138.

[20] Andrew J Tattersall and Penelope S Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (1996), 740–748.

[21] Christoph Tremmel, Christian Herff, Tetsuya Sato, Krzysztof Rechowicz, Yusuke Yamani, and Dean J Krusienski. 2019. Estimating cognitive workload in an interactive virtual reality environment using EEG. *Frontiers in human neuroscience* 13 (2019), 401.