Measuring Cognitive Load in Virtual Reality Training via Pupillometry

Joy Yeonjoo Lee, Nynke de Jong, Jeroen Donkers, Halszka Jarodzka, and Jeroen J. G. van Merriënboer

***Abstract*—Pupillometry is known as a reliable technique to measure cognitive load in learning and performance. However, its applicability to virtual reality (VR) environments, an emerging technology for simulation-based training, has not been well-verified in educational contexts. Specifically, the VR display causes light reflexes that confound task-evoked pupillary responses (TEPRs), impairing cognitive load measures. Through this pilot study, we validated whether task difficulty can predict cognitive load as measured by TEPRs corrected for the light reflex and if these TEPRs correlate with cognitive load self-ratings and performance. A total number of 14 students in health sciences performed observation tasks in two conditions: difficult versus easy tasks while watching a VR scenario in home health care. Then, a cognitive load self-rating ensued. We used a VR system with a built-in eye tracker and a photosensor installed to assess pupil diameter and light intensity during the scenario. Employing a method from the human–computer interaction field, we determined TEPRs by modeling the pupil light reflexes using a baseline. As predicted, the difficult task caused significantly larger TEPRs than the easy task. Only in the difficult task condition did TEPRs positively correlate with the performance measure. These results suggest that TEPRs are valid measures of cognitive load in VR training when corrected for the light reflex. It opens up possibilities of using real-time cognitive load for assessment and instructional design for VR training. Future studies should test our findings with a larger sample size, in various domains, involving complex VR functions such as haptic interaction.**

***Index Terms*—Cognitive load, educational simulations, medical training, mobile and personal devices, personalized e-learning, virtual and augmented reality.**

Manuscript received 4 April 2021; revised 1 August 2023 and 2 October 2023; accepted 18 October 2023. Date of publication 23 October 2023; date of current version 5 January 2024. This work was supported by The Netherlands Organization for Scientific Research under Grant 055.16.117. *(Corresponding author: Joy Yeonjoo Lee.)*

This work involves human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by FHML Research Ethics Committee under Application No. FHML-REC/2020/074/Amendment 1.

Joy Yeonjoo Lee is with the Faculty of Governance and Global Affairs, Leiden University, 2501 EE The Hague, The Netherlands (e-mail: j.y.lee@luc.leidenuniv.nl).

Nynke de Jong is with the Health Services Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands (e-mail: n.dejong@maastrichtuniversity.nl).

Jeroen Donkers and Jeroen J. G. van Merriënboer are with the School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands (e-mail: jeroen.donkers@maastrichtuniversity.nl; j.vanmerrienboer@maastrichtuniversity.nl).

Halszka Jarodzka is with the Faculty of Education Sciences, Open University, 6401 DL Heerlen, The Netherlands (e-mail: halszka.jarodzka@ou.nl).

Digital Object Identifier 10.1109/TLT.2023.3326473

I. Introduction

VIRTUAL reality (VR) has become a powerful alternative to physical simulation-based training [1]. VR creates immersive task environments that provoke a sense of presence, emotions, and engagement [2], [3], which allows for a favorable training environment with high fidelity of specific tasks [4], [5], [6]. Some studies have reported positive effects of VR training on learning outcomes and perceived usability [7], [8]. However, key challenges still remain: instructional design or course structure for VR training has not been fully established [6], [9], [10], and few studies have actually measured students' performance during VR training [7]. In order to build an effective instructional design for a new task environment, performance assessment and progress monitoring are fundamental [11]. This necessitates the development of real-time indicators of learning progress and task performance, which can be informed by cognitive load.

A. Cognitive Load Theory

Cognitive load provides a useful indicator to analyze learning processes in simulation training [12], [13], [14], [15]. Cognitive load theory posits an "element interactivity" where heterogeneous processes in cognitive, affective, and social domains coincide in working memory [13], [16], which is particularly relevant to simulation training that requires multiple tasks to be performed simultaneously within complex environments [12], [17], [18]. Cognitive load refers to the imposition of these processes caused by given tasks to working memory that has limited capacity [13], [19].

Cognitive load is not always harmful to task performance. There are different types of cognitive load that can be beneficial or detrimental, depending on the sources of cognitive load. In the traditional framework, three types of cognitive load are identified: intrinsic cognitive load that reflects the complexity of a task and a learner's competency for performing the task, germane cognitive load that pertains to learning, and extraneous cognitive load that stems from the suboptimal instructional design [13], [20]. For instance, if learning processes are motivated by working memory, it can be beneficial to performance but lead to a higher germane load. If the total cognitive load exceeds the working memory capacity, performance may deteriorate. In a recent framework of cognitive load, other types of cognitive load are identified: primary load for domain-specific task performance, and secondary load for domain-general metacognitive performance that supports the primary performance [21]. If the secondary load is activated, performance can be enhanced but

at the cost of a higher total cognitive load [22]. Accordingly, the literature has shown that the correlation between cognitive load and performance can either be positive or negative, depending on the characteristics of the measurements or the research context [12], [21], [22], [23].

If measured properly, cognitive load can be an effective indicator of performance, learning, and expertise, which in turn informs instructional design [13], [24]. Thus, measuring cognitive load with a valid and reliable method has been an important issue in research on learning and instruction [13], [25]. In general, three methods have been used to measure cognitive load: self-rating, secondary tasks, and psychophysiological indices [12], [25]. In simulation training, psychophysiological indices are shown to be the most sensitive measure of the three [12]. Specifically, eye-tracking indices have been used as a cognitive load measure for decades [26], [27], [28] demonstrating higher validity among other psychophysiological measures, such as heart rate or heart rate variability [12].

B. Using Pupillometry in Virtual Reality

Pupil dilation is a well-validated real-time measure of cognitive load in simulation training [24]. A large body of literature has confirmed that pupil dilation correlates with cognitive demands imposed by tasks, such as solving arithmetic problems or spelling difficult words [29], [30], [31], [32], [33]. At a physiological level, pupil dilation is known to be an involuntary response that reflects noradrenergic activity in the locus coeruleus, which regulates arousal, mental activity, and emotion [34], [35]. Since pupil dilation reflects emotional activity, it can be an effective measure of the cognitive load caused by task difficulty, especially when the task includes the management of emotion [23]. Moreover, pupil dilation may capture real-time changes in cognitive load more robustly than self-rating in dynamic environments such as computer-based simulation [23].

Nonetheless, when using pupillometry to measure cognitive load in dynamic task environments, researchers should be wary of a major confounder, the pupillary light reflex. The pupil dilation caused by cognitive processing for the given task, or task-evoked pupillary responses (TEPR), is notably small while the magnitude of change caused by light reflex is large [36]. A traditional way to control for this artifact is to obtain a baseline prior to the actual tasks, and calculate the difference between the baseline and the pupil size measured during the tasks [30], [37]. This baseline must be recorded in the same lighting conditions as the actual tasks and be developed for each participant as pupil diameter is highly idiosyncratic [38]. In VR environments with a head-mounted display (HMD), the most common approach is to use this method with fixed light conditions or fixed targets [39], [40].

However, when using real-world VR scenarios, using fixed lighting or targets is not practical or plausible. An alternative can be to develop baseline formulas that compute pupil dilation depending on the changing level of luminance, and calculate the difference between the baseline and measured pupil size (i.e., TEPR = pupil measured – pupil baseline). To quantify the luminance level, some studies used a photosensor implemented in the HMD [41], [42] while some estimated luminance by using color values of the pixels presented on the HMD [43], [44]. To find the baseline formulas, various mathematical models can be fitted depending on the experimental setups [45], [46], [47], [48], [49]. The subtractive correction of TEPR is applicable as the TEPR is reportedly independent of the baseline pupil size [50], [51].

Research on the use of pupillometry to measure cognitive states in VR is still in its infancy, and studies using light-reflex correction in VR are scarce. Moreover, most of the existing studies are in the field of engineering or human–computer interaction that uses simple cognitive tasks rather than complex real-life tasks for education. In a few studies that used a learning paradigm, the learning tasks were not sufficiently structured nor well-presented [52]. Among the studies measuring cognitive load in VR, the majority used a 2-D computer screen rather than HMD [52].

C. Present Study

This study aims to demonstrate TEPRs as an effective real-time measure of cognitive load in VR with HMD by using real-life problem solving for the education of healthcare professionals. First, TEPRs are determined by employing a correction method from the human–computer interaction field [41], [42], [53]. They then are validated as a cognitive load measure through two approaches: predictive validity that examines whether task difficulty as a factor of cognitive load predicts changes in TEPRs, and concurrent validity to test whether TEPRs positively correlate with other cognitive load measures such as self-rating. For this determination and validation, we test a hypothesis through an experimental setup: **H1**. Difficult tasks (DTs) evoke a higher cognitive load (measured by TEPRs) than easy tasks (ETs) in a VR training environment.

From an educational perspective, cognitive load should be interpreted in relation to performance [13]. Assuming the validity found from testing H1, we further explore correlations between cognitive load and performance in VR. According to the literature on cognitive load, the direction of the correlation can differ depending on task contexts and the sources of cognitive load. As our task deals with complex learning that requires diverse skills in real life (e.g., cognitive and metacognitive skills for patient safety), we do not specify the direction of correlation a priori, resulting in the second hypothesis: **H2**. The level of cognitive load correlates with the performance level in a VR training environment.

II. Methods

A. Participants and Design

A total of 14 undergraduate students (12 females; mean age 20.5; $SD = 1.5$) in Health Sciences were recruited at Maastricht University, The Netherlands. We used a within-subjects design with task difficulty as the single factor. Two different conditions were presented with the presentation order counterbalanced: ET condition where participants performed a simple observation task while watching a VR scenario, and DT condition with high task complexity.

B. Materials and Technical Setup

A VR scenario for home health care was developed by a research team at Maastricht University. In this scenario, a healthcare provider visits a patient to deliver medical care and share social interaction. This scenario takes 9 min and is formatted for a 360° HMD.

Two experts in healthcare professions education designed the observation tasks and the assessment. In ET, a simple instruction was given: "Observe and report: What is the homecare provider doing in the scenario?" In DT, more detailed observations were requested: "What are the patient's symptoms? Describe at least three symptoms; What are the strengths and weaknesses of the provider's performance? Describe at least three for each strengths and weaknesses." The observation results were reported in writing. The two experts composed a checklist consisting 7 to 15 items extracted from the CanMEDS model [54], the internationally recognized protocol for the competency of healthcare professionals. These items included professional roles (e.g., medical expertise, communication), patient information (e.g., age, past medical history), diagnosis (e.g., breathing, abnormality), and intervention (e.g., medication, social interaction) [55], [56]. They are selected as the most relevant parameters for the given scenario. The experts, then, assessed the accuracy of participants' reports based on the list.

A personal computer (Intel Core i9-9900K, 32 GB RAM) ran the scenario, displaying it through an HTC Vive Pro Eye headset (2880 × 1600 pixels, 110° visual field). This headset has a built-in eye tracker with 120 Hz sampling rate that uses HTC SRanipal SDK as an interface (version 1.3.1.0[1]). To quantify the luminance level, we implemented a photosensor (LDR sensor Iduino ST1107) in the HMD to import the luminance data. An Arduino board (Arduino UNO Rev3) connected the sensor to the computer via USB (see Fig. 1). The WorldViz Vizard software (version 6.0[2]) was used to arrange stimuli presentation and data recording. We used the VR Eye-tracking Analytics Lab package to synchronize the data from both the eye tracker and the photosensor.

For baseline recording, we used a separate VR scene where an empty room is presented. In this scene, we recorded participants' individual light reflexes for different light intensities [41], [57]. The light intensity was arranged to increase in ten stages, where each stage takes 5 s [38]. This scene does not include any additional visual or auditorial effects to form a neutral stimulus [44]. For the self-rating of cognitive load, we employed the widely used nine-point Paas Scale with the value 1 representing the lowest mental effort and the value 9 representing the highest [58].

C. Procedure

The participants were asked to read an informed consent form and sign it. Individual sessions consisted of two trials: ET and DT in counterbalanced order. Initially, participants were presented with instructions about the task. Next, they were positioned to stand at the center of the VR area, and then put on the VR headset. After a five-point calibration and the baseline session, the scenario was presented. Participants could rotate their heads or walk around in the VR area during the scenario. After the scenario, they reported their task results and self-rated their cognitive load for the task on the Paas Scale.

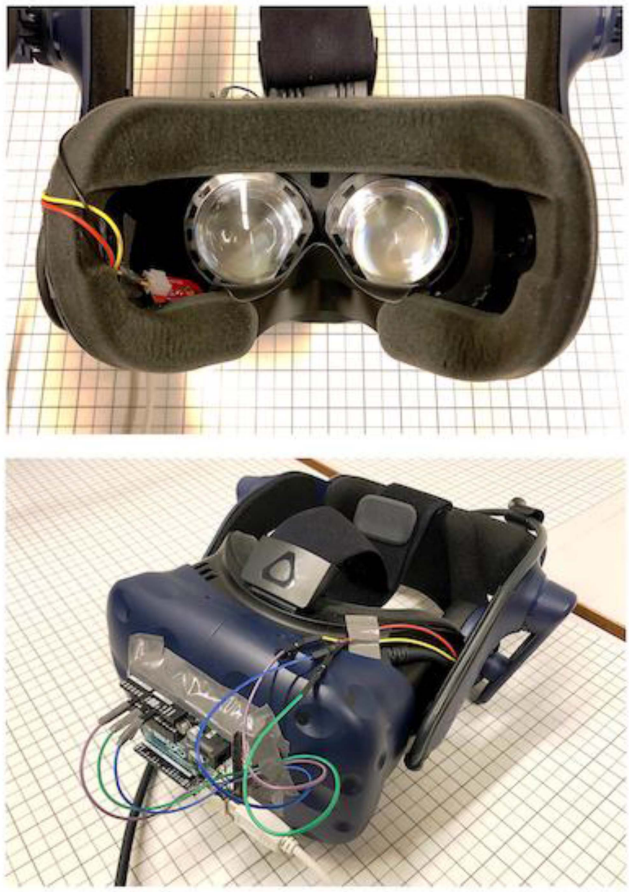

Fig. 1. On the upper side, a photosensor was installed inside the headset to measure the light intensity of the display. On the lower side, the Arduino board was mounted on the outside of the headset to connect the sensor to the computer via USB.

D. Data Analysis

The raw data included timestamps, light intensity, and pupil diameter. Using the baseline data, we developed models that represent the relationship between pupil diameter and light intensity. In the literature, various mathematical models of this relationship have been suggested for different contexts [45], [46], [47], [48], [49]. After visual inspection for initial data analysis, we applied an exponential model using least squares optimization. It is one of the most universal models [44], [49], with reportedly the lowest error in VR setups [59]. The model that fits all trials was sufficient to show that the method can function properly (F mean = 125.21, SD = 82.30; p mean = 0.00, SD = 0.00). Applying this model, pupil dilation caused by light reflexes was predicted to form a pupil baseline. The subtractive correction (i.e., TEPR = pupil measured – pupil baseline) was

[1][Online]. Available: www.vive.com
[2][Online]. Available: www.worldviz.com

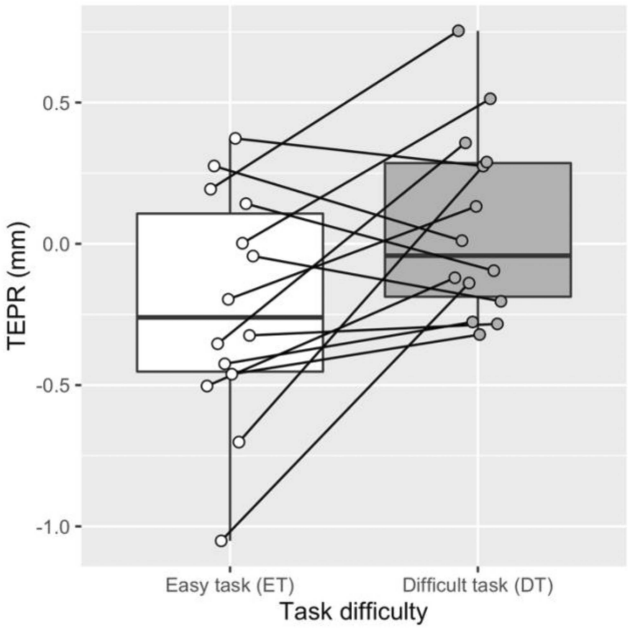

Fig. 2. Changes in TEPRs between ET and DT for all participants. The boxplots depict the quartile-based distribution of individual samples in each condition.

applied, resulting in TEPRs [50], [51]. The baseline pupil size before the correction was not significantly different between the two task conditions. The TEPRs then were averaged over time for each trial. The difference between ET and DT was tested via a dependent sample t test.

For calculating correlations, the two conditions of ET and DT were averaged for each participant, except when focusing on each condition. Based on the normality test of datasets, we used Spearman's method to calculate correlations. The task scores were averaged between the two raters. The inter-rater reliability was examined ($r_s = 0.89$, $p = 0.00$). R (version 3.5.1, R Core Team, 2019) was used for statistical analysis. We considered $p < 0.05$ to be statistically significant.

III. Results

A. TEPRs as a Cognitive Load Measure

Fig. 2 demonstrates the change in TEPRs between ET and DT for all 14 participants. As predicted, TEPRs were significantly larger in DT (mean = 0.07, $SD = 0.31$) than in ET (mean = −0.21, $SD = 0.43$) ($p = 0.01$). There was no significant difference in eye fixation position between ET and DT. The Paas Scale has significantly correlated with TEPRs ($r_s = 0.65$, $p = 0.00$). The self-rating score was higher in DT (median = 5, range = 2–7) than in ET (median = 4, range = 2–6), yet the difference between the conditions was not statistically significant ($p = 0.07$).

B. Correlations Between TEPRs and Performance

When averaging ET and DT, TEPRs did not significantly correlate with performance. However, when focusing on either condition, DT showed significant correlations. In DT, the task score was positively correlated with TEPRs ($r_s = 0.55, p = 0.04$) and the Paas Scale ($r_s = 0.68$, $p = 0.01$) (see Table I).

TABLE I
Correlations Between TEPRs, Paas Scale, and Task Score in DT Condition

Variable	TEPR	Paas Scale
Paas Scale	**0.54 (0.04)** [0.02,0.83]	
Task score	**0.55 (0.04)** [0.03,0.84]	**0.68 (0.01)** [0.24,0.89]

p values are presented in parentheses. Significant effects ($p < 0.05$) are in boldface. Values in square brackets indicate the 95% confidence interval for each correlation.

IV. Discussion

The present study tested the validity of TEPRs as a measure of cognitive load in a VR training environment. We have developed a VR learning environment where task complexity can be adjusted, established a VR system with an eye tracker, and controlled the light reflex to determine TEPRs. We have confirmed the predictive validity as cognitive load increased with task difficulty, and the concurrent validity as the cognitive load correlated with the self-ratings. In addition, performance measures correlated with TEPRs only in DT.

The first hypothesis (H1) assumed that cognitive load is higher in DTs than in ETs in a VR training environment. We found a significant impact of task complexity on cognitive load by using TEPRs, whereas the cognitive load self-rating did not show statistically significant effects. An explanation for the higher sensitivity of TEPRs compared with the self-rating might reside in the unique characteristics of VR task environments that provoke dynamic emotions through immersion [41]. While self-rating scales have been developed for classroom-based settings and depend solely on participants' judgment, pupil dilation is an involuntary response that reflects arousal and emotions [60]. Simulation training often involves affective and social tasks that require the management of emotions. Our observation tasks also included detecting the healthcare provider's empathy for the patient. In such task environments, pupillometry could be a more sensitive measurement of cognitive load than self-rating. This finding is consistent with previous research in simulation-based training [12], [23]. We suggest that more future research is needed to study the effect of emotional engagement on cognitive load in VR training.

Consequently, we largely confirm H1 based on the higher sensitivity of TEPRs as a cognitive load measure in VR training environments. This supports TEPR's predictive validity as a cognitive load measure. Although the cognitive load self-rating showed lower sensitivity, it significantly correlated with TEPRs,

which endorses TEPR's concurrent validity as a cognitive load measure.

The second hypothesis (H2) posited that the level of cognitive load correlates with performance level in a VR training environment. We found that cognitive load measured by the two indices (e.g., TEPRs and the Paas scale) positively correlates with the performance measure but only in the DT condition. Here, we suspect there is a ceiling effect in ET that might have caused weak discrimination in performance measures. We recommend that future studies should include a task analysis beforehand in order to prevent scale attenuation effects.

Concerning the positive direction of the correlation, it is likely that the sources of cognitive load in this study were rather beneficial to performance. For instance, they might have been stimulating information processing rather than harming performance. It appears that the task environment in this study is not as overwhelming as other simulation environments. The scenario was slow-paced, and the observation tasks were passive without any psychomotor skills required. Studies have shown that correlations between cognitive load and performance varied from positive to negative across different research contexts [12]. While this inconsistency may partly stem from measurement limitations [12], we argue that the nature of the factors that caused cognitive load determines the direction of the correlations. If the factors are negative to performance (e.g., distraction) or make the total cognitive load exceed working memory capacity, the cognitive load should be inversely proportional to performance. If the factors are positive (e.g., germane cognitive load [13], [61], self-regulation [23], [62]), cognitive load can positively correlate with performance. Again, we emphasize the importance of preceding task analysis to define the sources of cognitive load in future studies.

The validity and utility of TEPRs found by the present study open up new possibilities to improve research on VR training and instructional design for VR environments. Researchers may expand this finding to more diverse VR training environments, investigate the potential of pupillometry to assess cognitive processes during VR training, and test if this new measure can be used to evaluate performers' expertise. For educators who are searching for an effective assessment tool for VR training, TEPRs can be a good option that provides an objective indicator of performers' competence to manage situational and emotional challenges in complex environments. This assessment might compensate for the lacking discriminatory power of traditional performance measures (e.g., questionnaires, checklists), improving instructional design and training programs for VR environments.

Our study has several limitations. First, as a pilot study, we used a small sample of participants. This might reduce the generalizability of our findings. Second, the scenario included only one domain, i.e., home health care. Our findings should be tested if they are applicable to other domains. We recommend future studies to do task analysis for the given domain before the testing, so the measures can be properly operationalized for targeted constructs [11]. Third, the VR content we used was formatted for a 360° HMD, which is only one type of VR technology. Future studies should examine our methods in more advanced settings, such as 3-D-rendered VR with haptic interaction. For these studies, we propose a careful control for confounding factors as various sensory modalities are involved in such complex environments. Also, in different setups, methods to control for light reflex should be carefully chosen (e.g., using the same luminous level, photo sensors, or calculation of pixels on HMD). Finally, other confounding factors in pupillometry such as pupil foreshortening error (i.e., the influence of gaze position on pupil size) [60], [63] were not corrected, due to a lack of corresponding methods for VR environments. In the absence of such methods, we recommend making the experimental conditions comparable as much as possible (e.g., minimizing the difference in gaze position across the conditions).

To our knowledge, this study is the first to show the validity of TEPRs as a cognitive load measure in VR healthcare training. The hidden potential of using VR training lies in the utility of datasets from diverse sources such as eye tracking, which provides rich information about training development. Continued study is needed to improve the understanding of these datasets and make VR healthcare training more effective.

Acknowledgment

The authors would like to thank Dr. S. Metzelthin for her assistance with the performance assessment.

References

[1] J. O. Woolliscroft, "Innovation in response to the COVID-19 pandemic crisis," *Acad. Med.*, vol. 95, no. 8, pp. 1140–1142, 2020, doi: 10.1097/acm.0000000000003402.

[2] M. Slater, "Measuring presence: A response to the Witmer and Singer presence questionnaire," *Presence*, vol. 8, no. 5, pp. 560–565, 1999.

[3] G. Riva et al., "Affective interactions using virtual reality: The link between presence and emotions," *Cyberpsychol. Behav.*, vol. 10, no. 1, pp. 45–56, 2007, doi: 10.1089/cpb.2006.9993.

[4] N. Pellas, I. Kazanidis, and G. Palaigeorgiou, "A systematic literature review of mixed reality environments in K-12 education," *Educ. Inf. Technol.*, vol. 25, pp. 2481–2520, 2020.

[5] R. Hite et al., "Investigating potential relationships between adolescents' cognitive development and perceptions of presence in 3-D, haptic-enabled, virtual reality science instruction," *J. Sci. Educ. Technol.*, vol. 28, no. 3, pp. 265–284, 2019.

[6] Z. Merchant, E. T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, and T. J. Davis, "Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis," *Comput. Educ.*, vol. 70, pp. 29–40, 2014, doi: 10.1016/j.compedu.2013.07.033.

[7] N. Pellas, A. Dengel, and A. Christopoulos, "A scoping review of immersive virtual reality in STEM education," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 748–761, Oct./Dec. 2020.

[8] P. Pantelidis et al., "Virtual and augmented reality in medical education," in *Medical and Surgical Education: Past, Present and Future*, G. Tsoulfas, Ed. London, U.K.: IntechOpen, 2018, pp. 77–97.

[9] L. Jensen and F. Konradsen, "A review of the use of virtual reality head-mounted displays in education and training," *Educ. Inf. Technol.*, vol. 23, no. 4, pp. 1515–1529, 2018.

[10] C. Fertleman et al., "A discussion of virtual reality as a new tool for training healthcare professionals," *Front. Public Health*, vol. 6, 2018, Art. no. 44, doi: 10.3389/fpubh.2018.00044.

[11] J. J. Van Merriënboer and P. A. Kirschner, *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design*. Evanston, IL, USA: Routledge, 2018.

[12] L. M. Naismith and R. B. Cavalcanti, "Validity of cognitive load measures in simulation-based training: A systematic review," *Acad. Med.*, vol. 90, no. 11, pp. S24–S35, 2015, doi: 10.1097/ACM.0000000000000893.

[13] J. Sweller, J. J. Van Merriënboer, and F. Paas, "Cognitive architecture and instructional design: 20 years later," *Educ. Psychol. Rev.*, vol. 31, no. 2, pp. 261–292, 2019, doi: 10.1007/s10648-019-09465-5.

[14] K. L. Fraser, P. Ayres, and J. Sweller, "Cognitive load theory for the design of medical simulations," *Simul. Healthcare*, vol. 10, no. 5, pp. 295–307, 2015, doi: 10.1097/SIH.0000000000000097.
[15] R. Moreno and B. Park, "Cognitive load theory: Historical development and relation to other theories," in *Cognitive Load Theory*, J. L. Plass, R. Moreno, and R. Brünken, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2010, pp. 9–28.
[16] J. Q. Young, P. S. O'Sullivan, V. Ruddick, D. M. Irby, and O. T. Cate, "Improving handoffs curricula: Instructional techniques from cognitive load theory," *Acad. Med.*, vol. 92, no. 5, 2017, Art. no. 719, doi: 10.1097/acm.0000000000001664.
[17] F. A. Haji, D. Rojas, R. Childs, S. De Ribaupierre, and A. Dubrowski, "Measuring cognitive load: Performance, mental effort and simulation task complexity," *Med. Educ.*, vol. 49, no. 8, pp. 815–827, 2015, doi: 10.1111/medu.12773.
[18] J. Q. Young, S. M. Van Dijk, P. S. O'Sullivan, E. J. Custers, D. M. Irby, and O. Ten Cate, "Influence of learner knowledge and case complexity on handover accuracy and cognitive load: Results from a simulation study," *Med. Educ.*, vol. 50, no. 9, pp. 969–978, 2016, doi: 10.1111/medu.13107.
[19] J. J. Van Merrienboer and J. Sweller, "Cognitive load theory and complex learning: Recent developments and future directions," *Educ. Psychol. Rev.*, vol. 17, no. 2, pp. 147–177, 2005.
[20] J. Sweller, "Element interactivity and intrinsic, extraneous, and germane cognitive load," *Educ. Psychol. Rev.*, vol. 22, no. 2, pp. 123–138, 2010.
[21] J. Y. Lee, A. Szulewski, J. Q. Young, J. Donkers, H. Jarodzka, and J. J. G. Van Merriënboer, "The medical pause: Importance, processes, and training," *Med. Educ.*, vol. 55, pp. 1152–1160, 2021, doi: 10.1111/medu.14529.
[22] T. Seufert, "The interplay between self-regulation in learning and cognitive load," *Educ. Res. Rev.*, vol. 24, pp. 116–129, Jun. 2018, doi: 10.1016/j.edurev.2018.03.004.
[23] J. Y. Lee, J. Donkers, H. Jarodzka, G. Sellenraad, and J. J. G. Van Merriënboer, "Different effects of pausing on cognitive load in a medical simulation game," *Comput. Hum. Behav.*, vol. 110, 2020, Art. no. 106385, doi: 10.1016/j.chb.2020.106385.
[24] A. Szulewski, N. Roth, and D. Howes, "The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: A new tool for the assessment of expertise," *Acad. Med.*, vol. 90, no. 7, pp. 981–987, Jul. 2015, doi: 10.1097/Acm.0000000000000677.
[25] A. Korbach, R. Brünken, and B. Park, "Differentiating different types of cognitive load: A comparison of different measures," *Educ. Psychol. Rev.*, vol. 30, no. 2, pp. 1–27, 2018, doi: 10.1007/s10648-017-9404-8.
[26] M. G. Glaholt, "Eye tracking in the cockpit: A review of the relationships between eye movements and the aviators cognitive state," *Defence Res. Develop. Toronto (Canada)*, 2014.
[27] J. L. Rosch and J. J. Vogel-Walcutt, "A review of eye-tracking applications as tools for training," *Cogn. Technol. Work*, vol. 15, no. 3, pp. 313–327, 2013, doi: 10.1007/s10111-012-0234-7.
[28] B. A. Wilbanks and S. P. McMullan, "A review of measuring the cognitive workload of electronic health records," *CIN, Comput., Inform., Nurs.*, vol. 36, no. 12, pp. 579–588, 2018.
[29] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.
[30] J. Hyönä, J. Tommola, and A.-M. Alaja, "Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks," *Quart. J. Exp. Psychol.*, vol. 48, no. 3, pp. 598–612, 1995.
[31] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.
[32] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Hum. Factors*, vol. 43, no. 1, pp. 111–121, 2001.
[33] J. Klingner, B. Tversky, and P. Hanrahan, "Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks," *Psychophysiology*, vol. 48, no. 3, pp. 323–332, 2011, doi: 10.1111/j.1469-8986.2010.01069.x.
[34] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?," *Dev. Cogn. Neurosci.*, vol. 25, pp. 69–91, 2017, doi: 10.1016/j.dcn.2016.11.001.
[35] J. Brisson, M. Mainville, D. Mailloux, C. Beaulieu, J. Serres, and S. Sirois, "Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1322–1331, 2013, doi: 10.3758/s13428-013-0327-0.
[36] J. Beatty and B. Lucero-Wagoner, "The pupillary system," in *Handbook of Psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2000, ch. 6, pp. 142–162.
[37] P. W. M. Van Gerven, F. Paas, J. J. G. Van Merrienboer, and H. G. Schmidt, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, no. 2, pp. 167–174, 2004, doi: 10.1111/j.1469-8986.2003.00148.x.
[38] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. London, U.K.: Oxford Univ. Press, 2011.
[39] P. Bækgaard, J. P. Hansen, K. Minakata, and I. S. MacKenzie, "A Fitts' law study of pupil dilations in a head-mounted display," in *Proc. 11th ACM Symp. Eye Tracking Res. Appl.*, 2019, Paper 32, doi: 10.1145/3314111.3319831.
[40] C. Hirt, M. Eckard, and A. Kunz, "Stress generation and non-intrusive measurement in virtual environments using eye tracking," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 12, pp. 5977–5989, 2020, doi: 10.1007/s12652-020-01845-y.
[41] H. Chen, A. Dey, M. Billinghurst, and R. W. Lindeman, "Exploring pupil dilation in emotional virtual reality environments," in *Proc. Int. Conf. Artif. Reality Telexistence Eurograph. Symp. Virtual Environ.*, 2017.
[42] J. Iskander et al., "Exploring the effect of virtual depth on pupil diameter," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2019, pp. 1849–1854, doi: 10.1109/smc.2019.8913975.
[43] M. Eckert, E. A. P. Habets, and O. S. Rummukainen, "Cognitive load estimation based on pupillometry in virtual reality with uncontrolled scene lighting," in *Proc. 13th Int. Conf. Qual. Multimedia Experience*, 2021, pp. 73–76, doi: 10.1109/qomex51781.2021.9465417.
[44] M. Eckert, T. Robotham, E. A. P. Habets, and O. S. Rummukainen, "Pupillary light reflex correction for robust pupillometry in virtual reality," *Proc. ACM Comput. Graph. Interactive Techn.*, vol. 5, no. 2, pp. 1–16, 2022, doi: 10.1145/3530798.
[45] L. L. Holladay, "The fundamentals of glare and visibility," *J. Opt. Soc. Am.*, vol. 12, no. 4, pp. 271–319, 1926, doi: 10.1364/JOSA.12.000271.
[46] S. G. de Groot and J. W. Gebhard, "Pupil size as determined by adapting luminance∗," *J. Opt. Soc. Am.*, vol. 42, no. 7, pp. 492–495, 1952, doi: 10.1364/JOSA.42.000492.
[47] P. A. Stanley and A. K. Davies, "The effect of field of view size on steady-state pupil diameter," *Ophthalmic Physiol. Opt.*, vol. 15, no. 6, pp. 601–603, 1995, doi: 10.1016/0275-5408(94)00019-V.
[48] B. Winn, D. Whitaker, D. B. Elliott, and N. J. Phillips, "Factors affecting light-adapted pupil size in normal human subjects," *Invest. Ophthalmol. Vis. Sci.*, vol. 35, no. 3, pp. 1132–1137, 1994.
[49] A. B. Watson and J. I. Yellott, "A unified formula for light-adapted pupil size," *J. Vis.*, vol. 12, no. 10, pp. 12–12, 2012, doi: 10.1167/12.10.12.
[50] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychol. Bull.*, vol. 91, no. 2, 1982, Art. no. 276.
[51] J. Reilly, A. Kelly, S. H. Kim, S. Jett, and B. Zuckerman, "The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry," *Behav. Res. Methods*, vol. 51, no. 2, pp. 865–878, 2019, doi: 10.3758/s13428-018-1134-4.
[52] A. D. Souchet, S. Philippe, D. Lourdeaux, and L. Leroy, "Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review," *Int. J. Hum.–Comput. Interaction*, vol. 38, no. 9, pp. 801–824, 2022, doi: 10.1080/10447318.2021.1976509.
[53] B. Cebeci, U. Celikcan, and T. K. Capin, "A comprehensive study of the affective and physiological responses induced by dynamic virtual reality environments," *Comput. Animation Virtual Worlds*, vol. 30, no. 3/4, 2019, Art. no. e1893, doi: 10.1002/cav.1893.
[54] J. R. Frank and D. Danoff, "The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies," *Med. Teacher*, vol. 29, no. 7, pp. 642–647, 2007, doi: 10.1080/01421590701746983.
[55] J. Sherbino, K. Kulasegaram, A. Worster, and G. R. Norman, "The reliability of encounter cards to assess the CanMEDS roles," *Adv. Health Sci. Educ.*, vol. 18, no. 5, pp. 987–996, 2013, doi: 10.1007/s10459-012-9440-6.
[56] S. Chou, G. Cole, K. McLaughlin, and J. Lockyer, "CanMEDS evaluation in Canadian postgraduate training programmes: Tools used and programme director satisfaction," *Med. Educ.*, vol. 42, no. 9, pp. 879–886, 2008, doi: 10.1111/j.1365-2923.2008.03111.x.
[57] O. Palinko and A. L. Kun, "Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators," in *Proc. Symp. Eye Tracking Res. Appl.*, 2012, pp. 413–416, doi: 10.1145/2168556.2168650.

[58] F. G. W. C. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics—A cognitive-load approach," *J. Educ. Psychol.*, vol. 84, no. 4, pp. 429–434, Dec. 1992, doi: 10.1037/0022-0663.84.4.429.

[59] B. John, P. Raiturkar, A. Banerjee, and E. Jain, "An evaluation of pupillary light response models for 2D screens and VR HMDs," in *Proc. 24th ACM Symp. Virtual Reality Softw. Technol.*, 2018, Paper 19, doi: 10.1145/3281505.3281538.

[60] K. Holmqvist and R. Andersson, *Eye-Tracking: A Comprehensive Guide to Methods, Paradigms and Measures*. Lund, Sweden: Lund Eye-Tracking Research Inst., 2017.

[61] A. Szulewski, D. Howes, J. J. G. Van Merriënboer, and J. Sweller, "From theory to practice: The application of cognitive load theory to the practice of medicine," *Acad. Med.*, vol. 96, pp. 24–30, 2020, doi: 10.1097/acm.0000000000003524.

[62] A. B. H. De Bruin and J. J. G. Van Merriënboer, "Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research," *Learn. Instruct.*, vol. 51, pp. 1–9, 2017, doi: 10.1016/j.learninstruc.2017.06.001.

[63] T. R. Hayes and A. A. Petrov, "Mapping and correcting the influence of gaze position on pupil size measurements," *Behav. Res. Methods*, vol. 48, no. 2, pp. 510–527, Jun. 2016, doi: 10.3758/s13428-015-0588-x.

Joy Yeonjoo Lee received the Ph.D. degree (cum laude) from the School of Health Professions Education, Maastricht University, Maastricht, The Netherlands, in 2022.

She is an Assistant Professor of instructional technology and health data science with Leiden University, The Hague, The Netherlands. Her research utilizes eye tracking and data science to investigate human cognition and technology-enhanced learning based on educational theories (e.g., the Medical Pause, cognitive load). In particular, she has an active interest in using eye tracking in VR/AR environments and the application of AI to performance assessment and learning analytics.

Dr. Lee was the recipient of the SHE Dissertation Award 2022 form the School of Health Professions Education, Maastricht University.

Nynke de Jong received the master's degree in health sciences and the Ph.D. degree from Maastricht University, Maastricht, The Netherlands, in 2007. She finished her studies in nursing.

She is currently an Associate Professor with the Department of Health Services Research and the School of Health Professions Education, Maastricht University. Her research focuses on e-reality education.

Jeroen Donkers received the Ph.D. degree in artificial intelligence from Maastricht University, Maastricht, The Netherlands, in 2003.

He is a Assistant Professor with the Department of Educational Development and Research, as well as with the School of Health Professions Education, Maastricht University. He focusses his activities on smart use of computers in education for learning and assessment.

Halszka Jarodzka received the Diploma in psychology and the Ph.D. degree in pedagogical and media psycholgogy (dr.rer.nat.) from Eberhard-Karls University, Tuebingen, Germany, in 2007 and 2011, respectively.

She is currently a Full Professor of online learning and instruction from Open University, Heerlen, The Netherlands. Her research focuses on the use of eye-tracking in education to study and to foster learning, testing, and expertise development.

Jeroen J. G. van Merriënboer received the master's degree in experimental psychology and the Ph.D. degree in educational sciences from the University of Twente, The Netherlands, in 1990.

He is a Full Professor of learning and instruction with the Department of Educational Development and Research, as well as with the School of Health Professions Education, Maastricht University, Maastricht, The Netherlands. His research focuses on instructional design, specifically, cognitive load theory, four-component instructional design (4C/ID), and learning in the health professions.