Assessing Cognitive Load via Pupillometry

Pavel Weber(✉)[1], Franca Rupprecht[2], Stefan Wiesen[1], Bernd Hamann[3], and Achim Ebert[1]

[1] Department of Computer Science, University of Kaiserslautern, Germany
{weber,wiesen,ebert}@cs.uni-kl.de
[2] insight.out GmbH - digital diagnostics, Kaiserslautern, Germany
franca.rupprecht@insio.de
[3] Department of Computer Science, University of California, Davis, CA, U.S.A.
hamann@cs.ucdavis.edu

Abstract. A fierce search is called for a reliable, non-intrusive, and real-time capable method for assessing a person's experienced cognitive load. Software systems capable of adapting their complexity to the mental demand of their users would be beneficial in a variety of domains. The only disclosed algorithm that seems to reliably detect cognitive load in pupillometry signals – the Index of Pupillary Activity (IPA) – has not yet been sufficiently validated. We take a first step in validating the IPA by applying it to a working memory experiment with finely granulated levels of difficulty, and comparing the results to traditional pupillometry metrics analyzed in cognitive resarch. Our findings confirm the significant positive correlation between task difficulty and IPA the authors stated.

Keywords: Cognitive Load · Pupillometry · Index of Pupillary Activity (IPA) · Eye Tracking · Working Memory

Regular Research Paper

1 Introduction

Cognitive Load, understood as the amount of working memory resources dedicated to a specific task, determines a person's problem solving ability in terms of effectiveness and efficiency [18]. Best task performance is achieved based on a balanced, productive cognitive load level that avoids mental "under-challenge and over-challenge". A software system that is able to dynamically adapt task difficulty based on a person's experienced cognitive load in real time can have great impact on a variety of applications, e.g. learning, driving, and high-performance working environments such as that of pilots.

Many methods have been used to measure cognitive load, such as subjective self-reported measures and analytical approaches [7,2], and objective psychophysiological measures, e.g. electroencephalography (EEG), functional magnetic resonance imaging (fMRI), heart rate, blood pressure, skin temperature, and eye activity [21]. Many of these methods have the disadvantage of being intrusive

and depending on non-portable equipment. Thus have been tested only in controlled laboratory settings, and they require complex data analysis. Considering these limitations and the increasing accuracy and affordability of eye tracking systems, analysis of pupillometry data for extracting cognitive features has become increasingly feasible and common. The possibility of turning a smartphone, a tablet, or a webcam into an eye tracker emphasizes its real-word and real-time applicability [11,17].

While task-evoked pupillary response (TEPR) was found to be a reliable measure directly corresponding to working memory [6], it does not distinguish between pupillary reflex reactions to light changes and reactions induced by cognitive effort. The only published algorithm that claimed to successfully separate light reflexes from dilation reflexes is the index of pupillary activity (IPA). The IPA was published broadly and openly, in contrast to the patented index of cognitive activity (ICA) that was used in a wide range of studies.

We present a validation of the IPA by applying it to an experiment with finely granulated levels of difficulty of cognitive tasks, and compare the results with traditional TEPR metrics, namely the percentage change of pupil diameter (PCPD). First, we analyze a participant's performance for the different levels of task difficulty. Specific hypotheses and expectations have guided our efforts: We expect reaction times to increase and accuracy to decrease with increasing difficulty. Next, we calculate the PCPD and analyze its peaks and magnitudes. We expect both the peaks and magnitudes to increase with increasing difficulty. Finally, we calculate the IPA for each trial, expecting it to also increase with increasing difficulty level. In conclusion we discuss limitations of the proposed method and provide incentives for future work.

2 Background and Related Work

2.1 TEPR

The correlation between pupil diameter and problem difficulty has already been noted in the 60s [8]. In short-term memory tasks it was observed that the pupil dilated during the presentation phase, and constricted during the recall phase. The peak pupil diameters were found to be directly related to the number of items presented [10]. Other studies found the raw pupil diameter to be not comparable across participants and proposed the PCPD as a metric of interest [12,9]. It is computed in regard of a certain baseline, typically an average value over a given amount of seconds of pupil diameter data measured before the experiment.

The main problem with this measures is that the changes in pupil size for the most part cannot definitely be attributed to either lighting conditions or actual cognitive effort. It was found that changes in pupil diameter size evoked by light reflexes can be described as large (up to a few millimeters), while those evoked by cognitive activity happen to be relatively small (usually between 0.1 and 0.5 mm) and rapid [1]. Those are however very loose ranges and cannot be directly applied to reliably distinguish the cause of the pupillary reflex.

2.2 ICA and IPA

In the early 2000s Marshall developed the ICA which seems to be able to distinguish the pupillary reflexes [13]. The ICA uses wavelet analysis to compute the rate of occurences of abrupt discontinuities in the pupil diameter signal. The assumption is that low IPA values (i.e. few abrupt discontinuities per time period) reflect little cognitive effort while high values indicate strong cognitive effort. Although the algorithm itself is proprietary and its implementation is undisclosed, it has been used in a variety of studies and is claimed to be reliable across sampling rates and different hardware platforms [20,4,3].

Since there is no independent verification of the ICA, another research group has developed their own version of the algorithm, using clues in different papers and the patent manuscript: the IPA [5]. The IPA uses discrete wavelet transformation – similar to the ICA – but differs in the choice of wavelet, thresholding approach, and extrema detection method. The algorithm itself is disclosed in the paper, making it possible for other researchers to reproduce every step of it. A multi-level wavelet decomposition with a *Symlet-16* mother-wavelet is used to separate low-frequency components (level-1 detail coefficients), corresponding to light reflexes, and high-frequency components (level-2 detail coefficients), triggered by cognitive activity. The modulus maxima is used to find local extrema in the level-2 coefficients. Those maxima are then compared to a so-called universal threshold, denoted by $\sigma\sqrt{(2\log n)}$. All maxima above this threshold are considered an abrupt discontinuity.

3 Method

3.1 Study Design

The present study was a within-subjects eye tracking experiment based on a simple Memory Span Task. Memory Span Tasks are used to determine a user's Working Memory Capacity (WMC). With each new trial, the participant is presented an increasing number of items and then asked to recall them. The WMC is the longest number of sequential items that the user can correctly recall. For a typical young adult the WMC is 7 ± 2 [15]. In our adapted version the number of digits was not successively increased but randomized. We were not so much interested in the participant's WMC but rather in inducing different levels of intrinsic cognitive load that somewhat reflect non-ideal real-world conditions.

The independent variable is the number of digits presented—which represents the inherent task difficulty for the trial. The dependent variables are reaction time, answer correctness, and the pupil diameters measured during the whole experiment. From the signal of the latter we calculated PCPD as well as IPA values for each trial.

3.2 Participants

The study was conducted as part of a HCI lecture with a sample of 34 international students. Data of 4 participants had to be discarded due to difficulties

with eye tracker calibration, giving a final sample size of $N = 30$ (16 female, 14 male) with age ranging from 22 to 45 ($\mu = 26 \pm 4$).

3.3 Apparatus

The laboratory was setup in a clean and neutral office with two windows covered by roller blind, and two double flux fluorescent ceiling lights. The inside lighting condition fluctuated between 160 and 192 Lux during the experiment period of two weeks, depending on the weather. A PupilLabs Pupil Core eye tracking system was used to acquire pupillometry data for both eyes at a rate of 120 Hz. The corresponding Pupil Capture software as well as a specially developed JavaFX application were executed on a customary windows machine with a 1080p monitor and standard mouse and keyboard. Brightness and contrast of the display were constant.

3.4 Procedure

After filling out a simple demographic questionnaire, the participant was asked to put on the head-mounted eye tracker. We then started a marker-based calibration sequence since it yielded the best confidence values of the system. Once the confidence level was stable and high enough, the actual Digit Span Task was started. Instead of being asked to recall the digits, the participant was shown a composite number sequence with the same number of digits. This composite sequence could contain built-in errors, hence the participant was asked whether it corresponds exactly to the single digits that were shown before. The answer was given by pressing the left arrow key for *no*, the right arrow key for *yes* respectively. The participant had the opportunity to get acquainted with the task by performing six training trials with 3 to 5 digits. The training was followed by five blocks of 24 trials each, with sequences of 3 to 10 digits, resulting in a total of 120 trials per participant (15 for each difficulty level). Each block took around 3.6 ± 0.3 minutes to complete. The participants were encouraged to take a short break after each block.

4 Results

4.1 Analyses

Since the relationship of the difficulty levels cannot be assumed to be linear, but definitely to be monotonic, Spearman's rank coefficient ρ was used to calculate correlation. In addition, repeated measures of analyses of variance (ANOVA) were used to find significant effects. Cohen's parameter d was calculated to assess the effect sizes and variation between group means, emphasizing respective significance through pairwise T-tests. Descriptive statistics for all dependent variables are summarized in Table 1 while effect size and significance are shown in Table 2. All training data was excluded from the analyses. All analyses were conducted in Python, more specifically *Pandas* and *Pingouin*.

4.2 Task Performance

Task performance was measured in terms of reaction time and answer correctness. Reaction time was expected to increase with increased task difficulty, while correctness was expected to decrease with increased task difficulty, i.e., the presumption was that more difficult tasks would take more time to complete and were more likely to be answered wrongly. Our analyses confirmed this presumption. Figure 1 shows a clear linear trend for both observations.

The correlation between task difficulty and reaction time was moderately positive $(0.54, p < .001)$ and the one between task difficulty and answer correctness was weakly negative $(-0.22, p < .001)$.

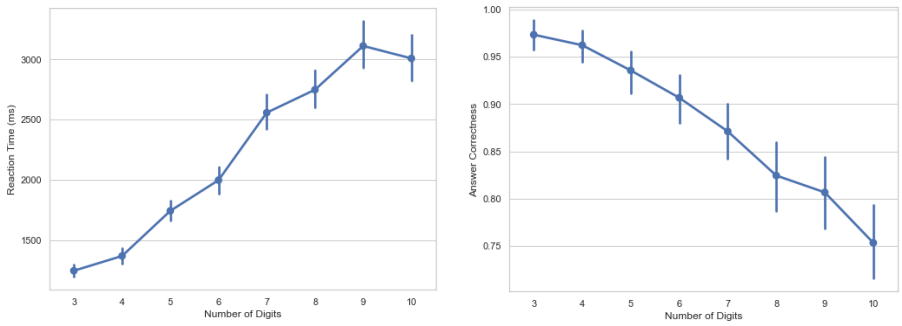

Fig. 1: Average reaction times and answer correctness per number of digits.

ANOVA revealed significant effects of both: task difficulty having impact on reaction time, $F_{(7,203)} = 69.62, p < .001, \eta^2 = 0.706$, and on correctness, $F_{(7,203)} = 23.67, p < .001, \eta^2 = 0.449$. This proves both our hypotheses regarding task performance, i.e., reaction time increases and accuracy decreases with increasing difficulty.

Table 2 shows that the effects of task difficulty levels on reaction time are highly significant, except for the differences between 7 and 8 digits (being significant) and between 9 and 10 digits (not being significant). Concerning answer correctness, the difference between 9 and 10 digits is the only highly significant one. All others are significant, except for the differences between 8 and 9 digits and between 3 and 4 digits. Even though the effect sizes are not large, these results support our paradigm of performing testing with finely granulated task difficulty levels.

4.3 Pupil Dilation

We used the median pupil diameter of the training sequence as baseline for calculating the PCPD. PCPD values for each trial were aggregated to the actual variables of interest, i.e., pupil dilation peaks and magnitudes between pupil

Table 1: Statistic results for dependent variables of the experiment; effect of different task difficulty levels.

digits	**reaction time** μ and σ (s)	**accuracy** μ and σ (%)	**PCPD peak**, μ and σ (%)	**magnitude** μ and σ (%)	**IPA** μ and σ (Hz)
3	1.25 ± 0.60	97.3 ± 16.1	$.033 \pm .118$	$.129 \pm .058$	$1.134 \pm .378$
4	1.37 ± 0.70	96.2 ± 19.1	$.034 \pm .116$	$.134 \pm .067$	$1.155 \pm .350$
5	1.74 ± 0.92	93.5 ± 24.6	$.039 \pm .115$	$.142 \pm .065$	$1.199 \pm .360$
6	2.00 ± 1.23	90.6 ± 29.1	$.036 \pm .133$	$.151 \pm .098$	$1.183 \pm .351$
7	2.56 ± 1.53	87.1 ± 33.5	$.046 \pm .128$	$.159 \pm .081$	$1.216 \pm .346$
8	2.75 ± 1.75	82.4 ± 38.1	$.042 \pm .116$	$.154 \pm .061$	$1.239 \pm .352$
9	3.11 ± 2.18	80.6 ± 39.5	$.049 \pm .125$	$.167 \pm .077$	$1.222 \pm .331$
10	3.00 ± 2.06	75.3 ± 43.1	$.048 \pm .121$	$.168 \pm .074$	$1.236 \pm .336$

Table 2: Effect sizes and significance between task difficulty level groups.

pair	**reaction time** d	p	**accuracy** d	p	**PCPD peak**, d	p	**magnitude** d	p	**IPA** d	p
(3,4)	.074	***	-.034	-	.009	-	.062	-	.061	-
(4,5)	.229	***	-.082	**	.044	-	.101	**	.126	**
(5,6)	.156	***	-.088	**	-.021	-	.126	**	-.049	-
(6,7)	.343	***	-.111	**	.076	-	.109	*	.096	*
(7,8)	.114	**	-.138	**	-.031	-	-.072	*	-.065	-
(8,9)	.225	***	-.053	-	.062	-	.168	***	-.050	-
(9,10)	-.064	-	-.169	***	-.015	-	.008	-	.040	-

Significance: *$p < .05$, **$p < .01$, ***$p < .001$

dilation valleys and peaks. Figure 2 illustrates the average values for the two metrics considered. The trends of both graphs match our expectations.

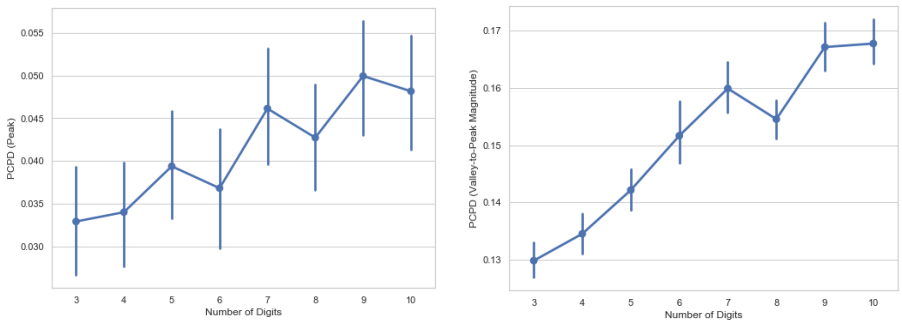

Fig. 2: Average PCPD peaks and valley-to-peak magnitudes per number of digits.

The correlation between PCPD peaks and task difficulty was weakly positive $(0.059, p < .001)$, with a significant effect $F_{(7,203)} = 4.144, p < .001, \eta^2 = 0.125$. This proves our first TEPR-related hypothesis that PCPD peaks increase with task difficulty. However, pairwise T-tests revealed that none of the between-groups effects were significant.

The second correlation of interest is the correlation between PCPD valley-to-peak magnitudes and task difficulty. This one is stronger, but it is also only weakly positive $(0.225, p < .001)$. ANOVA revealed a significant effect: $F_{(7,203)} = 22.645, p < .001, \eta^2 = 0.438$. This result proves our second TEPR-related hypothesis that PCPD valley-to-peak magnitudes increase with task difficulty. Regarding the effects between difficulty level groups, 5 out of 7 pairs showed a significant effect of at least $p < .05$, see Table 2. The only pairs that showed no significant effect were between 3 and 4 digits and between 9 and 10 digits.

4.4 Abrupt Discontinuities

Our next goal was aimed at determining whether the same behavior holds for the more sensitive metric that is said to distinguish between pupil dilation reflex and light reflex, the IPA. We calculated IPA values for every commenced second of a trial and averaged values.

Figure 3 shows that the correlation between IPA and task difficulty is indeed weakly positive $(0.105, p < .001)$ with a significant effect $(F_{7,203} = 16.327, p < .001, \eta^2 = 0.36)$. Table 2 reveals, however, that only two effects of difficulty level differences are significant, the one between 4 and 5 digits and the one between 6 and 7 digits with the latter only being $p < .05$. Nevertheless, the general significance for the effect of task difficulty on IPA, shown by the ANOVA, is not questioned.

The second plot in Figure 3 shows another interesting detail about the IPA. Following cognitive load theory, the highest experienced cognitive load should occur during the stimulus presentation phase and the lowest during pauses, while the mental load experienced during the recall phase should not have a direct correlation to the task difficulty but should follow the same trend as the task inherent difficulty, therefore the mental load should be low for simple tasks and increase significantly with higher difficulty levels [19]. This behavior — together with the weak positive correlation between IPA and PCPD magnitude $(0.112, p < .001)$ — confirms our last hypothesis that the IPA is an indicator for cognitive load that increases with task difficulty.

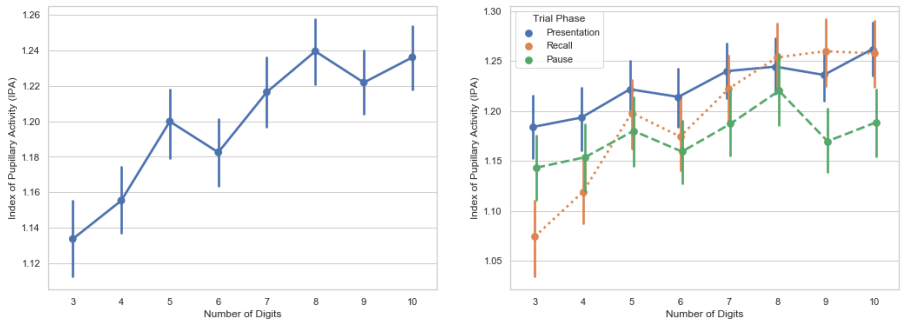

Fig. 3: Average IPA per number of digits and trial phase.

5 Conclusions

The results of our statistical analyses validate our specific hypotheses, as stated in the Introduction. We have shown that finely granulated difficulty settings can have significant impact on task performance. We have shown that impact on task performance is reflected in the magnitude of pupil dilation amplitudes. Finally, we have ascertained that this behavior is also substantiated by the values calculated with the relatively new IPA algorithm. Our experiments have demonstrated that the IPA correlates with traditional TEPR metrics, even in finely granulated task difficulty settings. While the authors differentiated between three difficulty settings, and found no significant effect between the easy and the control tasks, we found two significant effects that occured when increasing the number sequence by just one digit, additionally to the general significant effect of task difficulty on IPA revealed by ANOVA. Effect size and significance levels for the different difficulty levels were not as high for measured IPA data relative to measured TEPR data. In summary, our findings validate the IPA. However, we found the unmodified IPA algorithm to be sensitive to sampling rate and signal length, resulting in very different recognized discontinuity counts. The chosen symlet-16

mother wavelet therefore seems not to be universally applicable, hence our IPA values differ from the ranges reported by the authors.

These findings underline the need for further investigation to ensure smooth utilization of the IPA. Our experimental design considered only one task, i.e., the digit span task. Since the digits were all shown at the center of the screen, we did not consider eye tracking measures such as fixation and saccade. We have not yet analyzed eye blink frequency and latency. The authors of the IPA removed all data points in a 200 ms window before the start and after the end of a detected blink. Our method, in contrast, uses cubic spline interpolation to reconstruct the signal (see [14]). While the cubic spline approach seems to work well, it would seem of interest to determine how well it performs in a more realistic experimental setting.

Concerning possible future research directions, it would certainly be useful to further examine the validity of the IPA through a battery of tests. Unlike the ICA there is the possibility to do so independently. Users of the IPA would benefit from a large-scale comparison of different sample rates, wavelets, and coefficient resolutions, possibly resulting in proper usage guidelines. The algorithm also remains to be tested under varying light conditions, like it was done for the ICA [16], as well as under different tasks and modalities. It may also be of interest how the measured cognitive load value relates to the relative performance of individuals. These various research directions must be supported by the results of other valid cognitive feature extraction methods.

Acknowledgements: This research was partially funded by the German research foundation (DFG) within the IRTG 2057 “Physical Modeling for Virtual Manufacturing Systems and Processes”.

References

1. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin **91**(2), 276–292 (1982). https://doi.org/10.1037/0033-2909.91.2.276
2. Cook, A., Zheng, R., Blaz, J.: Measurement of cognitive load during multimedia learning activities. Cognitive Effects of Multimedia Learning pp. 34–50 (Jan 2008). https://doi.org/10.4018/978-1-60566-158-2.ch003
3. Demberg, V.: Pupillometry: the index of cognitive activity in a dual-task study. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 35, pp. 2154–2159 (2013)
4. Demberg, V., Sayeed, A.: The Frequency of Rapid Pupil Dilations as a Measure of Linguistic Processing Difficulty. PLOS ONE **11**(1), e0146194 (Jan 2016). https://doi.org/10.1371/journal.pone.0146194
5. Duchowski, A.T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., Giannopoulos, I.: The Index of Pupillary Activity: Measuring Cognitive Load *vis-à-vis* Task Difficulty with Pupil Oscillation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18. pp. 1–13. ACM Press, Montreal QC, Canada (2018). https://doi.org/10.1145/3173574.3173856

6. Granholm, E., Asarnow, R.F., Sarkin, A.J., Dykes, K.L.: Pupillary responses index cognitive resource limitations. Psychophysiology **33**(4), 457–461 (1996). https://doi.org/10.1111/j.1469-8986.1996.tb01071.x
7. Hart, S.G.: Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **50**(9), 904–908 (Oct 2006). https://doi.org/10.1177/154193120605000909
8. Hess, E.H., Polt, J.M.: Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science **143**(3611), 1190–1192 (Mar 1964). https://doi.org/10.1126/science.143.3611.1190
9. Jiang, X., Atkins, M.S., Tien, G., Bednarik, R., Zheng, B.: Pupil responses during discrete goal-directed movements. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14. pp. 2075–2084. ACM Press, Toronto, Ontario, Canada (2014). https://doi.org/10.1145/2556288.2557086
10. Kahneman, D., Beatty, J.: Pupil Diameter and Load on Memory. Science **154**(3756), 1583–1585 (Dec 1966). https://doi.org/10.1126/science.154.3756.1583
11. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye Tracking for Everyone. arXiv:1606.05814 [cs] (Jun 2016)
12. Kruger, J.L., Steyn, F.: Subtitles and Eye Tracking: Reading and Performance. Reading Research Quarterly **49**(1), 105–120 (2014). https://doi.org/10.1002/rrq.59
13. Marshall, S.: The Index of Cognitive Activity: Measuring cognitive workload. In: Proceedings of the IEEE 7th Conference on Human Factors and Power Plants. pp. 7–5–7–9. IEEE, Scottsdale, AZ, USA (2002). https://doi.org/10.1109/HFPP.2002.1042860
14. Mathôt, S., Fabius, J., Van Heusden, E., Van der Stigchel, S.: Safe and sensible preprocessing and baseline correction of pupil-size data. Behavior Research Methods **50**(1), 94–106 (2018). https://doi.org/10.3758/s13428-017-1007-2
15. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review **63**(2), 81–97 (1956). https://doi.org/10.1037/h0043158
16. Rerhaye, L., Blaser, T., Alexander, T.: Evaluation of the index of cognitive activity (ica) as an instrument to measure cognitive workload under differing light conditions. In: Congress of the International Ergonomics Association. pp. 350–359. Springer (2018). https://doi.org/10.1007/978-3-319-96059-3_38
17. Semmelmann, K., Weigelt, S.: Online webcam-based eye tracking in cognitive science: A first look. Behavior Research Methods **50**(2), 451–465 (Apr 2018). https://doi.org/10.3758/s13428-017-0913-7
18. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Science **12**(2), 257–285 (Apr 1988). https://doi.org/10.1016/0364-0213(88)90023-7
19. Sweller, J.: Cognitive Load Theory: Recent Theoretical Advances (Apr 2010). https://doi.org/10.1017/CBO9780511844744.004
20. Vogels, J., Demberg, V., Kray, J.: The Index of Cognitive Activity as a Measure of Cognitive Processing Load in Dual Task Settings. Frontiers in Psychology **9**, 2276 (Nov 2018). https://doi.org/10.3389/fpsyg.2018.02276
21. Zheng, R.Z., Greenberg, K.: The boundary of different approaches in cognitive load measurementstrengths and limitations. In: Cognitive Load Measurement and Application, pp. 45–56. Routledge (2017). https://doi.org/10.4324/9781315296258-4