

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/363711456>

Cognitive Load Measurement Using Arithmetic and Graphical Tasks and Galvanic Skin Response

Chapter · September 2022

DOI: 10.1007/978-3-031-16014-1_66

CITATIONS

3

READS

17

3 authors:



Zihisire Muke Patient

Wroclaw University of Science and Technology

12 PUBLICATIONS 18 CITATIONS

[SEE PROFILE](#)



Zbigniew Telec

Wroclaw University of Science and Technology

49 PUBLICATIONS 498 CITATIONS

[SEE PROFILE](#)



Bogdan Trawinski

Wroclaw University of Science and Technology

134 PUBLICATIONS 1,178 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:






Mobile learning [View project](#)



impact of university websites on the enrollment of foreign students [View project](#)



Cognitive Load Measurement Using Arithmetic and Graphical Tasks and Galvanic Skin Response

Patient Zihisire Muke , Zbigniew Telec , and Bogdan Trawiński^(✉) 

Department of Applied Informatics, Wrocław University of Science and Technology, Wrocław,
Poland

{patient.zihisire, zbigniew.telec, bogdan.trawinski}@pwr.edu.pl

Abstract. The results of an experiment to measure cognitive load using arithmetic and graphical tasks and galvanic skin response (GSR) biometric technique are presented in this paper. 62 volunteers were recruited to take part in a laboratory experiment conducted with the integrated iMotions biometric platform. Data were collected using observations, Single Ease Question (SEQ) and NASA Task Load Index (NASA-TLX) self-report questionnaires, and GSR measurements. The 18 performance, subjective, and psychophysiological indicators were calculated from the collected data to measure the cognitive load associated with arithmetic and graphical tasks. Nonparametric tests of statistical significance of differences between individual metrics were made for the easy, medium, and hard arithmetic and graphical tasks. The conducted research proved the usefulness of most measures in the analysis of the cognitive load associated with arithmetic and graphical tasks.

Keywords: Cognitive load · Arithmetic tasks · Graphical tasks · Galvanic skin response · NASA-TLX questionnaire

1 Introduction

Cognitive load in cognitive psychology has been described as the load imposed on a person's working memory by a specific (learning) task [1]. In relation to HCI (Human Computer Interaction), cognitive load is described as the quantity of mental effort that is used from the working memory while performing cognitive tasks and interacting with the computer system [2]. Assessing the cognitive load could be useful in designing user interfaces and user experience evaluation. Data from biosensors could provide the possibility to qualify the behavioral responses caused by given software, insight into missing information, or lack of understanding. A significant number of diverse biometric techniques such as electroencephalography (EEG), eye tracking, facial expression analysis, galvanic skin response (GSR), electrocardiography (ECG) and electromyography (EMG), allow researchers to perform comparative analysis, which can determine the reflection of cognitive load levels [3].

The aim of the research presented in this paper was to explore cognitive load while participants completed arithmetic and graphical matrix reasoning tasks. A multimodal approach to measurement was applied which included performance measurements, subjective self-reports with Single Ease Question (SEQ) and NASA Task Load Index (NASA-TLX) and Single Ease Question (SEQ) questionnaires, and biometric measurements using several metrics based on psychophysiological data from the GSR sensor.

The iMotions biometric research platform [4] was used to conduct an experiment on 62 participants, mostly young people with a university education. The designed stimuli were consistent with standard tests used to measure cognitive workload. Arithmetic and graphical tasks of low, medium, and difficult difficulty were used as the stimuli. The arithmetic task involved mentally calculating mathematical expressions involving addition and subtraction of numbers with varying numbers of digits. Graphical tasks, on the other hand, were a type of graphical matrix reasoning in which the participant's task was to find a missing element in a matrix containing abstract shapes. The collected data was analyzed using non-parametric tests in order to determine the statistical significance of the differences between the individual measurements.

As part of this research, three Master's theses were completed and they contain a detailed description of the experiment presented in this paper [5–7]. Due to the limited space in the paper, we present here only a part of the results. The first part of the research on impact of the Stroop effect on cognitive load was presented by the authors in [8]. The experiment was conducted during the COVID-19 pandemic period, therefore, the participants and research team had to observe strict laboratory hygienic rules. The study was approved by the Research Ethics Committee at the Wroclaw University of Science and Technology, Poland.

2 Background and Related Works

2.1 Cognitive Load

The concept of cognitive load arises from the early work done in the area of education including teaching and was expressed in the cognitive load theory (CLT) designed in the late 1980s by the psychologist John Sweller to describe how human short and long-term memory works and how the human brain processes and stores information [9, 10].

CLT was constructed based on the human memory model that includes the sensory, working and long-term memory subsystems [11]. The sensory memory holds the correct sensory information of what is briefly presented (i.e., <0.25 s). They come from the five senses: smell, vision, taste, touch and hearing. On the other hand, the working memory is only able to process certain pieces of information and retains the more processed input information for a short period of time (i.e., <30 s). At last, the long-term memory is a repository of all the knowledge of a learner for a long period of time. This model prove that CLT can be utilized to limit unnecessary effort in this area [12].

To fully understand the theory of cognitive load, in connection with the instructional area, it is essential to classify knowledge in two categories. The first one is called the biologically primary knowledge and defined as the basic knowledge that can be

acquired naturally and effortlessly by living beings through evolution for many generations. Learning a first language as a child is a great example of biologically primary knowledge, and both speaking and hearing do not require teaching. The second one is named the biologically secondary knowledge and describe as the knowledge acquired through intellectual effort, which must be learned through explicit instructions. A good example of this are writing skills, which require awareness and demanding continuous work [13, 14].

Furthermore, in CLT, three types of cognitive load can be distinguished: intrinsic, extraneous and germane load. The intrinsic cognitive load refers to the inherent difficulty or complexity of a task or material. Activities vary depending on the level of difficulty. If the task is too complex or difficult, it overloads the working memory and makes it too tough to process. A great example is an equation containing derivatives or integrals which is more difficult to solve than the simple task of adding two numbers [15, 16].

The extraneous load is a subset of cognitive load which is generated by the presence of irrelevant information that distracts people from the normal learning process. For example, if a graphic and text are not properly combined, the learner will spend more time switching from graphics to text. This information overloads the working memory and produces extraneous cognitive load [15, 16].

The germane load, on the other hand, is defined as the “effective cognitive load”. This is the result of properly representing information in working memory and organizing it into schemas. For example, someone creates flowcharts that will be more helpful to learners than plain text with steps to be followed. Systematic ordering of information accelerates its memorization [15, 16].

2.2 Cognitive Load Measurements

Cognitive load can be measured using two approaches. The first is an analytical technique where the main emphasis is on assessing mental load and collecting analytical metrics based on tasks, as well as analyzing mathematical models and subjective measures, more specifically expert opinions. The second is the empirical technique that focuses on subjective measures, which include a self-rating approach, and objective measures, which consist of a performance or task based approach, physiological approach and behavioral approach. The empirical method is more often used by researchers than the analytical method [17].

Regarding the empirical technique, methods based on subjective evaluation rely on participants’ own judgment of their task performance efforts. The user evaluates his own cognitive processes and the mental effort required for a given task. In this method, there are two types of rating scales: the unidimensional scales, which measure the overall cognitive load, e.g. the Paas’ nine-point mental effort rating scale [18] and the multidimensional scales, which focus on different components of workload, e.g. one of the most commonly used is NASA-TLX (NASA-Task Load Index) with six dimensions: mental demand, physical demand, time pressure, mental effort, performance satisfaction, and frustration level [19]. Another used subjective rating scale measure especially in user experience testing and usability, is the Single Ease Question (SEQ). It is a 7-point rating scale to evaluate how difficult participants find a task [20]. NASA-TLX and SEQ were used in this study.

In terms of objective measures, the performance-based approach is often used in dual-task environments to measure a user's workload based on their own performance in secondary tasks compared to primary tasks, e.g. task completion times, completion rates, error rates and test scores [17]. In this study performance metrics were as well measured. On the other hand, the behavioral approach implicitly and objectively records the subjects' actions, e.g. eye activity like fixation duration and blink frequency, mouse usage, speech features and head movements [21].

Finally, the psychophysiological approach based on neuroscience is an effective strategy for measuring cognitive load with the advantages of real-time, objectivity and low data reading invasiveness. Examples are: EEG, ECG, EMG, GSR, facial expression analysis and eye tracking [3].

2.3 Cognitive Load and Arithmetic Tasks

Applying mental arithmetic tasks in research is quite common in the study of the mechanisms of human cognitive activity because of its complex mental function. Mental arithmetic appears to engage the memory processes in retrieving arithmetic facts from the long-term memory and can be considered as a standardized stress-inducing experimental protocol [22].

Thus, many researchers in the literature used arithmetic tasks to measure cognitive load. In [23] six metrics of GSR were studied in the perception of four levels of cognitive load with emotional interference. Data were derived from two arithmetic experiments and emotions were elicited by displaying unpleasant and pleasant images in the background. Two classifiers were used to detect the level of cognitive load. The results showed that the identified metrics were able to detect four and two levels of cognitive load with high accuracy, even in the case of emotional changes.

In addition, the study in [24] presented and evaluated several metrics from blink and GSR signals to classify levels of cognitive stress. The experiment included four levels of difficulty with the use of arithmetic tasks, and two types of machine learning algorithms were used for classification. The results obtained showed that the blink and GSR metrics tested could reasonably differentiate the levels of cognitive load, and the combination of the metrics could improve the accuracy of the cognitive load classification.

Moreover, in [25] participants were classified as bad and good counting based on their performance while performing mental arithmetic tasks. Mathematical calculations actively engaged participants in mental work. In this study, the mental load concept was used with the EEG to classify signals. As a result, the participants were successfully classified into classes with 80% accuracy.

2.4 Cognitive Load and Graphical Matrix Reasoning Tasks

The concept of matrix reasoning evokes the measurement of visual and fluid intelligence through a series of incomplete visual matrices. The participants of an experiment or test have to choose one pattern from the five suggested options that best complete the matrix [26].

Raven's Advanced Progressive Matrices is one of the most widely used measures of visual and fluid intelligence. This measure is a matrix reasoning test in which subjects

are given a matrix of 3×3 with patterns representing a figure lacking the bottom right pattern. The patterns form a figure, and participants have to choose the appropriate pattern to fill in the missing space with one of eight suggested options [27].

In [27] it was stated that the Raven's matrices is a task designed to answer the question of whether matrix reasoning problems requiring new rule combinations correspond more closely to fluid intelligence and working memory capacity than problems requiring repetitive rule combinations.

In the same perspective, [28] presents a study where a multimodal sensing approach was used (GSR, ECG, blood oxygen content saturation and respiration) and three cognitive tests were performed, consisting of a Raven's matrices reasoning test, a numerical test, and a video game to induce cognitive workload in 12 subjects on the easy and hard levels of task difficulty. This research presents preliminary results in which a set of cognitive workload indicators is identified.

Another measure of the working memory capacity is the Matrix Reasoning Item Bank (MaRs-IB) which contains graphical tasks very similar to Raven's matrices. A description of the MaRs-IB can be found in [29]. Our study employed the same type of matrix reasoning to access cognitive load.

2.5 Galvanic Skin Response

The galvanic skin response (GSR) is part of the electrodermal activity measurement that tracks the activity of the sweat glands. Its measures can be divided into two groups that are tonic and phasic. The skin's conductance level is a good example of tonic responses and is thought to reflect changes in arousal. The phasic reactions include the skin conductivity response, which is a change in electrical conductivity. Positive and negative stimuli can increase arousal. An increase in the level of excitation causes an increase in skin conductivity [30, 31].

An increase in the difficulty level of a task intensifies cognitive load. Previous studies show a link between stress and cognitive load. Therefore, the number of GSR peaks could also be a determinant of cognitive load, which tends to increase with the cognitive load level. Researchers list further GSR measures used to measure cognitive load such as accumulative and mean galvanic skin response. The accumulative GSR is the summation of GSR values over the time of the ongoing task, from the first occurrence of the stimulus till its end. On the other hand, the mean GSR is the value of the GSR data divided by the average from the entire session [32–34].

3 Experiment Setup

The experiment was conducted on the iMotions biometric platform [4] using a variety of stimuli and biometric techniques. Due to the limited space, in this paper we present only a part of the results concerning human cognitive load associated with arithmetic and graphical tasks using a single biosensor technique GSR along with other subjective and performance measurements.

Based on the literature, we formulated three main hypotheses. First, the values of the GSR psychophysiological metrics increase with increasing cognitive load levels in

both arithmetic and graphical tasks (*H1*). Second, the subjective ratings based on the NASA-TLX and SEQ self-report questionnaires increase with increasing cognitive load levels in both arithmetic and graphical tasks (*H2*). Third, the values of the performance metrics are good supplementary indicators of the level of cognitive load (*H3*).

3.1 Participants

62 volunteers took part in the experiment, including 14 females and 48 males. Most of the respondents had higher education, which guaranteed that they would be able to solve the tasks. 39 people were university students, 6 PhD students, 9 IT specialists, one high school student and 6 other professions. 53 were right-handed, 6 were left-handed, and 3 were without a dominant hand. 42 respondents had good eyesight, 15 wore glasses and 5 wore contact lenses. They were between the ages of 20 and 38. 26 participants came from Poland, 24 from India, 10 from Indonesia and 2 from other countries. Before starting the experiment, participants signed their consent to participate in the study. They were instructed on the course of the experiment.

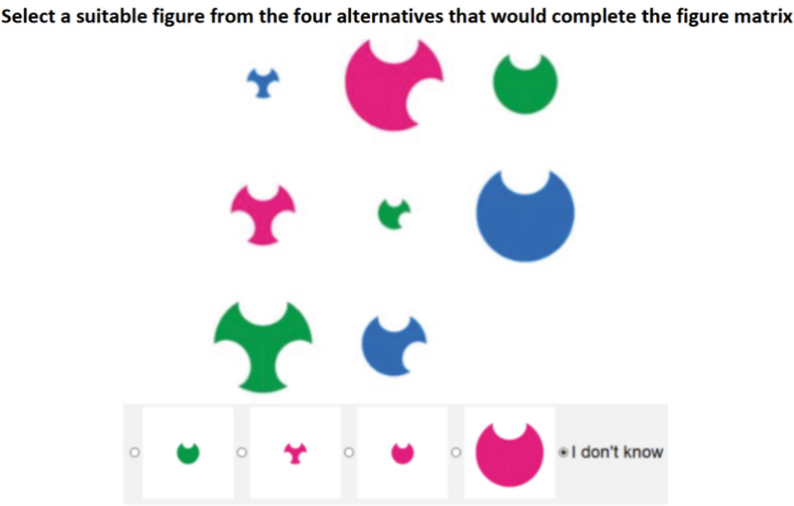
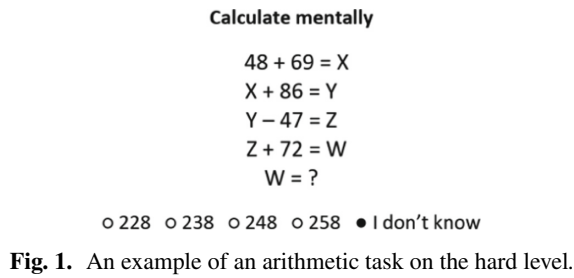
3.2 Arithmetic and Graphical Tasks

In our experiment, we used two types of stimuli in the form of arithmetic and graphical tasks. Each type consisted of three groups of five tasks at easy, medium and hard levels.

The arithmetic tasks consisted in mentally calculating a set of four equations with unknowns and determining the value of the last variable. The participant chose the correct answer from among four options. Each difficulty level differed in the size of numbers added and subtracted as well as partial results and the complexity of logical reasoning. At the easy level, the arguments were single-digit numbers and the partial scores were greater than zero. At the medium level, the arguments were two-digit numbers between 10 and 29 and the partial scores were greater than 10 and less than 100. On the hard level, on the other hand, the arguments were two-digit numbers between 41 and 99 and the calculated variables were larger than 100 and smaller than 300. Similar calculations were used in [35]. An example of an arithmetic task on the hard level is shown in Fig. 1. In the rest of the article, the arithmetic tasks are denoted as *ArL*, *ArM*, and *ArH*, respectively, for tasks on the low, medium, and hard difficulty levels.

Graphical tasks were developed based on the Matrix Reasoning Item Bank (MaRs-IB) [29]. The questions were in the form of a 3×3 matrix containing eight abstract graphics. The participants deduced relations between graphics that could differ in four dimensions: color, size, position and shape, and then selected the missing ninth graphics from four options. Difficulty levels were based on the dimensionality of individual tasks, which represented the number relation changes in the matrix. Tasks on the easy, medium, and hard levels had dimensions of 1, 2 and 3, and 4 to 6, respectively. An example of a graphical task on the medium level is presented in Fig. 2. In the further part of the paper, the graphical tasks are marked as *GrL*, *GrM* and *GrH*, which means tasks on the low, medium and hard difficulty level, respectively.

The time limit for solving arithmetic and graphical tasks on the easy, medium, and hard levels was 25, 40, and 50 s, respectively.



3.3 Measurement of Cognitive Load

For subjective measurement, we applied two types of self-report questionnaires that participants completed after finishing each group of tasks. A modified version of the NASA TLX questionnaire was used in this study because we did not use a weighting scheme in calculating the total score. Also, instead of the original scale, we used a 7-point Likert scale to score each question. We hypothesized that this scale might make it easier for participants to make a more reliable assessment in less time. The second questionnaire was the Single Ease Question (SEQ). The SEQ is a 7-point rating scale that was used to determine participants' perceived level of task difficulty.

A Shimmer GSR device was used to measure the electrodermal activity. The GSR electrodes were placed on the index and middle fingers of the participants' non-dominant hand. The participants used their dominant hand to operate the mouse.

In total, 18 performance, subjective and psychophysiological metrics were defined and calculated on the basis of the collected data. They are presented in Table 1.

Due to incorrect GSR measurements, we rejected the results of 5 participants. Consequently, all metrics were calculated from the data collected for 57 respondents. Moreover,

Table 1. Metrics used to measure cognitive load

Type	Metrics	Denotation
Performance metrics	Task completion rate	<i>Tcr</i>
	Task completion time	<i>Tct</i>
Subjective metrics	Task difficulty (SEQ)	<i>Seq</i>
	NASA-TLX - Overall	<i>NXo</i>
	NASA-TLX - Complexity	<i>NXc</i>
	NASA-TLX - Physical demand	<i>NXp</i>
	NASA-TLX - Time pressure	<i>NXt</i>
	NASA-TLX - Performance satisfaction	<i>NXs</i>
	NASA-TLX - Mental effort	<i>NXm</i>
	NASA-TLX - Frustration	<i>NXf</i>
Psychophysiological metrics	GSR Peak count	<i>Gpc</i>
	GSR Peaks per minute	<i>Gpm</i>
	GSR Mean value	<i>Gmv</i>
	GSR Mean value (normalized)	<i>Gmn</i>
	GSR Max value	<i>Gxv</i>
	GSR Max value (normalized)	<i>Gxn</i>
	GSR Accumulative value	<i>Gav</i>
	GSR Accumulative value (normalized)	<i>Gan</i>

taking into account that the participants performed five tasks on each difficulty level, the results are the average value of the measurements of the five tasks. Only, the task completion rate was calculated as the percentage of successfully completed tasks. Scores for each question from the SEQ and NASA-TLX questionnaires were a separate metric and they ranged from 1 to 7. The overall NASA-TLX score was calculated as the average of the scores for the six survey questions.

Psychophysiological metrics were calculated for the data collected from the GSR sensor using two algorithms implemented in the iMotions platform, namely GSR Peak Detection [36] and GSR Epoching Pre-processing [37]. The *Gmn*, *Gxn*, and *Gan* values were normalized for individual participants by dividing *Gmv*, *Gxv*, and *Gav* by the baseline values determined when filling in the personal questionnaires. A detailed description of the GSR metrics is contained in our previous paper presenting the first part of the results of our experiment [8].

4 Analysis of Experiment Results

The mean and median values of individual performance, subjective, and psychophysiological metrics are presented in Figs. 3, 4 and 5, respectively.

Prevailing majority of the results produced by the performance, subjective and psychophysiological metrics did not have a normal distribution according to the Shapiro-Wilk test. Therefore, the nonparametric Friedman test was applied, followed by the post-hoc Nemenyi test for multiple comparisons. The null hypotheses assumed that there were no significant differences in the values of individual metrics between the tasks. The significance level for rejecting the null hypothesis was set at 0.05. The results of the Nemenyi test for individual metrics are presented in Table 2, where the + sign means that the value of a given metric is statistically significantly larger for the task on the higher difficulty level compared to the task on the lower difficulty level. Sign – indicates that the value of the metric is statistically significantly smaller for the task on the higher difficulty level compared to the task on the lower difficulty level. The \approx sign, on the other hand, denotes that the null hypothesis was not rejected.

For the psychophysiological measurements the values of the *Gpc*, *Gav*, and *Gan* metrics for hard tasks (*ArH* and *GrH*) were significantly greater than the values for medium tasks (*ArM* and *GrM*), and these, in turn, were significantly greater than the values for easy tasks (*ArE* and *GrE*). The values of the *Gmv*, *Gmn*, *Gxv*, and *Gxn* metrics were consistent with this observation for arithmetic tasks except for *ArE* and *ArM*, between which no statistically significant differences occurred. Also, the *Gpm* metric values for graphical tasks were in line with the main observation. In turn, the results provided by the *Gmv*, *Gmn*, *Gxv*, and *Gxn* metrics for graphical tasks were inconsistent.

In the case of subjective measurements, the values of the seven *Seq*, *Nxo*, *NXc*, *NXp*, *NXt*, *NXm*, and *NXf* metrics for hard tasks (*ArH* and *GrH*) were significantly higher than the values for medium tasks (*ArM* and *GrM*), and these, in turn, were significantly greater than the values for easy tasks (*ArE* and *GrE*). Only for the *NXs* metrics there were no significant differences between all arithmetic and graphical tasks.

Among performance metrics, the task completion time *Tct* increased significantly with increasing difficulty of arithmetic and graphical tasks. In turn, the task completion rate *Tcr* significantly decreased with increasing difficulty of arithmetic and graphical tasks, except for arithmetic easy and medium tasks. There were no significant differences in *Tcr* between the *ArE* and *ArM* tasks.

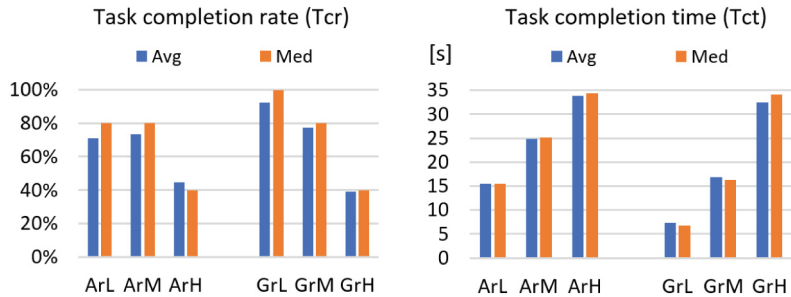


Fig. 3. Results of performance metrics

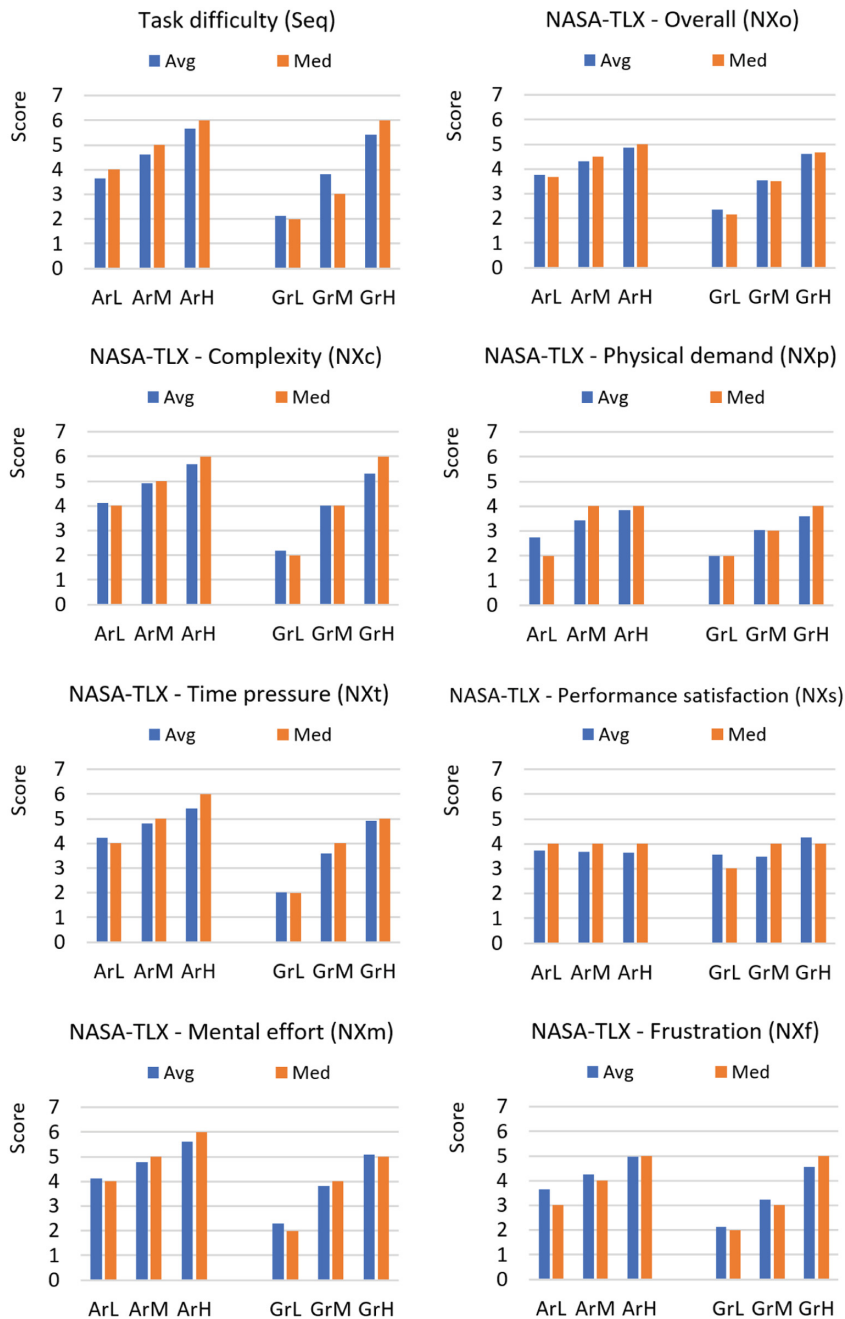


Fig. 4. Results of subjective metrics

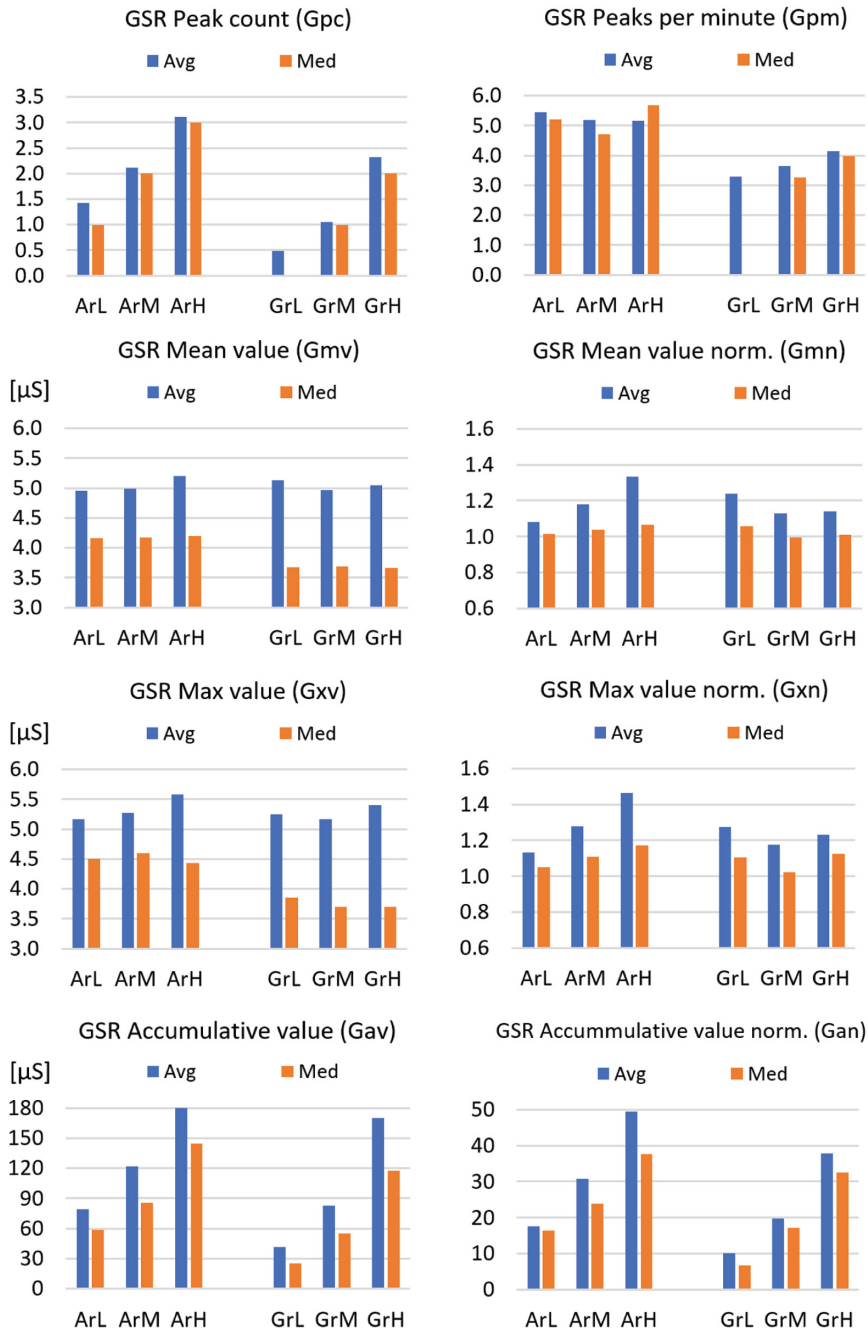


Fig. 5. Results of psychophysiological metrics

Table 2. Nemenyi test results for comparing the arithmetic and graphical tasks

Metrics	ArL vs ArM	ArL vs ArH	ArM vs ArH	Metrics	GrL vs GrM	GrL vs GrH	GrM vs GrH
<i>Tcr</i>	≈	—	—	<i>Tcr</i>	—	—	—
<i>Tct</i>	+	+	+	<i>Tct</i>	+	+	+
<i>Seq</i>	+	+	+	<i>Seq</i>	+	+	+
<i>NXo</i>	+	+	+	<i>NXo</i>	+	+	+
<i>NXc</i>	+	+	+	<i>NXc</i>	+	+	+
<i>NXp</i>	+	+	+	<i>NXp</i>	+	+	+
<i>NXt</i>	+	+	+	<i>NXt</i>	+	+	+
<i>NXs</i>	≈	≈	≈	<i>NXs</i>	≈	≈	≈
<i>NXm</i>	+	+	+	<i>NXm</i>	+	+	+
<i>NXf</i>	+	+	+	<i>NXf</i>	+	+	+
<i>Gpc</i>	+	+	+	<i>Gpc</i>	+	+	+
<i>Gpm</i>	≈	≈	≈	<i>Gpm</i>	≈	+	+
<i>Gmv</i>	≈	+	+	<i>Gmv</i>	—	≈	≈
<i>Gmn</i>	≈	+	+	<i>Gmn</i>	—	≈	≈
<i>Gxv</i>	≈	+	+	<i>Gxv</i>	—	≈	+
<i>Gxn</i>	+	+	+	<i>Gxn</i>	—	≈	+
<i>Gav</i>	+	+	+	<i>Gav</i>	+	+	+
<i>Gan</i>	+	+	+	<i>Gan</i>	+	+	+

5 Conclusions

The experiment in this study was conducted under laboratory conditions for one month with 62 participants who followed strict sanitary procedures. Age and education restrictions were also taken into account when recruiting participants, so that age and prior knowledge did not affect the results. The stimulus design was based on standardized tests that were validated to measure cognitive load. They consisted of 15 arithmetic and 15 graphical reasoning tasks of easy, medium and hard difficulty levels, with 5 tasks at each difficulty level. All tasks were implemented in a standalone application developed in Python. Biometric data collection was done using several biosensors from the iMotions platform. However, due to the limitation of the paper volume, among biometric data, only GSR metrics are presented, calculated with the use of algorithms built into the iMotions platform.

The collected biometric data was also normalized with baseline data in order to eliminate subjective dependencies. Apart from biometric data, subjective and performance measures were also analyzed. The calculated measures were visualized in graphs and tables to facilitate review of the experimental data and to compare the level of cognitive load based on the two designed stimuli. Moreover, statistical tests were performed to determine the impact of each stimulus on the measurements.

Based on statistical tests, the most common observations were as follows: the values of the metrics for the hard tasks (*ArH* and *GrH*) were significantly greater than the values for the medium tasks (*ArM* and *GrM*), and these in turn were significantly greater than the values for the easy tasks (*ArE* and *GrE*). This occurred for the performance metrics: *Tct*, subjective metrics: *Seq*, *Nxo*, *NXc*, *NXp*, *NXt*, *NXm*, and *NXf*, and psychophysiological

metrics: *Gpc*, *Gav*, and *Gan*. The task completion rate *Tcr* changed significantly in an inverse manner, except for the arithmetic tasks *ArE* and *ArM*, between which no significant differences were observed..

Regarding biometric data, eight GSR metrics were calculated to provide insight into the experienced level of cognitive load. The results show that the *H1* hypothesis is supported in the case of the *Gpc*, *Gav* and *Gan* metrics which increase when task difficulty is increasing for both arithmetical and graphical tasks. Whereas, for the *Gpm*, *Gmv*, *Gmn*, *Gxv* and *Gxn* metrics *H1* was not confirmed. The exceptions were *Gpm* metrics for graphical tasks as well as for *Gmn* and *Gxn* for arithmetic tasks, where the increase in GSR values with increasing difficulty levels occurred for either arithmetic tasks or for graphical tasks.

Concerning subjective measurements, eight metrics were analysed. The results indicate that *Seq*, *Nxo*, *NXc*, *NXp*, *NXt*, *NXm*, and *NXf* metrics support the *H2* hypothesis where the obtained values increased significantly with the difficulty of the tasks for both arithmetic and graphical tasks. In contrast, *H2* was not proved for the *NXs* performance satisfaction metrics.

In the case of performance measurements, two metrics were examined. The results show that *Tct* increased significantly with increasing difficulty of arithmetic and graphical tasks, thus supporting the *H3* hypothesis. *Tcr*, which changes in the opposite direction, i.e. decreases with the difficulty of the task, for graphical tasks also supports *H3*. *Tcr* did not reveal significant differences only for the *ArE* and *ArH* tasks. This indicates that *Tcr* requires larger differences in the difficulty levels for the arithmetic tasks to distinguish them.

In conclusion, our study demonstrated the usefulness of both arithmetic and graphical tasks as stimuli and the performance, subjective, and psychophysiological metrics we used to measure cognitive load. The following metrics are most appropriate for measuring cognitive load when participants perform arithmetic and graphical tasks: *Tct*, *Seq*, *Nxo*, *NXc*, *NXp*, *NXt*, *NXm*, *NXf*, *Gpc*, *Gav*, and *Gan*.

Other types of graphical reasoning tasks can be considered to ensure deep involvement in memory processes. In addition, a greater variety of difficulty levels of arithmetic tasks can be provided.

The future work will concentrate on including HCI stimuli in the study such as mobile and web applications along with different biometric techniques using multiple machine learning algorithms to predict the level of users' cognitive load from the human-computer interaction perspective.

References

1. Van Gog, T., Paas, F.: Cognitive load measurement. In: Seel, N.M. (ed.) *Encyclopedia of the Sciences of Learning*. Springer, Boston (2012). https://doi.org/10.1007/978-1-4419-1428-6_412
2. Kumar, N., Kumar, J.: Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Procedia Comput. Sci.* **84**, 70–78 (2016). <https://doi.org/10.1016/j.procs.2016.04.068>

3. Zihisire Muke, P., Trawinski, B.: Concept of research into cognitive load in human-computer interaction using biometric techniques. In: Proceedings of the PP-RAI 2019 Conference, Wrocław, Poland, pp. 78–83 (2019). http://pp-rai.pwr.edu.pl/PPRAI19_proceedings.pdf. Accessed 01 June 2022
4. iMotions Biometric Research Platform (8.1): iMotions A/S, Copenhagen, Denmark (2020)
5. Bresso, P.: Study of the impact of various stimuli on human cognitive load using electroencephalography and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław (2020)
6. Desai, H.: Study of the impact of various stimuli on human cognitive load using eye tracking and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław (2021)
7. Maharani, P.A.: Study of the impact of various stimuli on human cognitive load using facial expression analysis and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław (2020)
8. Zihisire Muke, P., Piwowarczyk, M., Telec, Z., Trawiński, B., Maharani, P.A., Bresso, P.: Impact of the Stroop effect on cognitive load using subjective and psychophysiological measures. In: Nguyen, N.T., Iliadis, L., Maglogiannis, I., Trawiński, B. (eds.) ICCCI 2021. LNCS (LNAI), vol. 12876, pp. 180–196. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88081-1_14
9. Sweller, J.: Cognitive load during problem solving: effects on learning. *Cogn. Sci.* **12**(1), 257–285 (1988). [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
10. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* **4**(4), 295–312 (1994). [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
11. Young, J.Q., Van Merriënboer, J., Durning, S., Ten Cate, O.: Cognitive Load Theory: Implications for medical education: AMEE Guide No. 86. *Med. Teach.* **36**(5), 371–384 (2014). <https://doi.org/10.3109/0142159X.2014.889290>
12. McLeod, S.A.: Multi store model of memory. Simply Psychology (2017). <https://www.simplypsychology.org/multi-store.html>
13. Sweller, J., Ayres, P., Kalyuga, S.: Cognitive Load Theory. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies, vol. 1. Springer, New York (2011). <https://doi.org/10.1007/978-1-4419-8126-4>
14. Geary, D.: An evolutionarily informed education science. *Educ. Psychol.* **43**(4), 179–195 (2008). <https://doi.org/10.1080/00461520802392133>
15. Orru, G., Longo, L.: The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2018. CCIS, vol. 1012, pp. 23–48. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14273-5_3
16. Sweller, J., van Merriënboer, J.J.G., Paas, F.: Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **31**(2), 261–292 (2019). <https://doi.org/10.1007/s10648-019-09465-5>
17. Paas, F., Tuovinen, J., Tabbers, H., Van Gerven, P.: Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* **38**(1), 63–71 (2003)
18. Paas, F.: Training strategies for attaining transfer of problem solving skills in statistics: a cognitive load approach. *J. Educ. Psychol.* **84**, 429–434 (1992)
19. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* **53**(1), 61–86 (2004). <https://doi.org/10.1111/j.1464-0597.2004.00161>
20. Gibson, A., et al.: Assessing usability testing for people living with dementia. In: REHAB 2016: Proceedings of the 4th Workshop on ICTs for improving Patients Rehabilitation Research Techniques, pp. 25–31 (2016). <https://doi.org/10.1145/3051488.3051492>

21. Chen, F., et al.: Robust Multimodal Cognitive Load Measurement. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-31700-7>
22. Zyma, I., et al.: Electroencephalograms during mental arithmetic task performance. *Data* **4**(1), 2–7 (2019). <https://doi.org/10.3390/data4010014>
23. Nourbakhsh, N., Chen, F., Wang, Y., Calvo, R.A.: Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Trans. Interact. Intell. Syst.* **7**(3), 1–12 (2017). <https://doi.org/10.1145/2960413>. Article 12
24. Nourbakhsh, N., Wang, Y., Chen, F.: GSR and blink features for cognitive load classification. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) *INTERACT 2013*. LNCS, vol. 8117, pp. 159–166. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40483-2_11
25. Rai, A.A., Ahirwal, M.K.: Electroencephalogram-based cognitive load classification during mental arithmetic task. In: Patgiri, R., Bandyopadhyay, S., Borah, M.D., Emilia Balas, V. (eds.) *Edge Analytics. LNEE*, vol. 869, pp. 479–487. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0019-8_36
26. Kievit, R.A., et al.: Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychol. Sci.* **28**(10), 1419–1431 (2017). <https://doi.org/10.1177/0956797617710785>
27. Harrison, T.L., Shipstead, Z., Engle, R.W.: Why is working memory capacity related to matrix reasoning tasks? *Mem. Cogn.* **43**(3), 389–396 (2014). <https://doi.org/10.3758/s13421-014-0473-3>
28. Hirachan, N., Mathews, A., Romero, J., Rojas, R.F.: Measuring cognitive workload using multimodal sensors, pp. 2–5 (2022). <http://arxiv.org/abs/2205.04235>
29. Chierchia, G., Fuhrmann, D., Knoll, L.J., Pi-Sunyer, B.P., Sakhardande, A.L., Blakemore, S.J.: The matrix reasoning item bank (MaRs-IB): novel, open-access abstract reasoning items for adolescents and adults. *Roy. Soc. Open Sci.* **6**(10), 1–13 (2019). <https://doi.org/10.1098/rsos.190232>
30. Braithwaite, J., Watson, D., Jones, R., Rowe, M.: A guide for analysing electrodermal activity (EDA) skin conductance responses (SCRs) for psychological experiments. Technical report, 2nd version. University of Birmingham, UK (2015)
31. Farnsworth, B.: What is GSR (galvanic skin response) and how does it work? (2018) <https://imotions.com/blog/gsr/>
32. Yoshihiro, S., Takumi, Y., Koji, S., Akinori, H., Koichi, I., Tetsuo, K.: Use of frequency domain analysis of skin conductance for evaluation of mental workload. *J. Physiol. Anthropol.* **27**(4), 173–177 (2008)
33. Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F.: Galvanic skin response (GSR) as an index of cognitive load. In: *CHI 2007 Extended Abstracts on Human Factors in Computing Systems*, pp. 2651–2656 (2007). <https://doi.org/10.1145/1240866.1241057>
34. Nourbakhsh, N., Wang, Y., Chen, F., Calvo, R.: Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: *24th Australian Computer-Human Interaction Conference (OzCHI)*, Melbourne, Australia, pp. 420–423. ACM Press (2012). <https://doi.org/10.1145/2414536.2414602>
35. Sinharay, A., Chatterjee, D., Sinha, A.: Evaluation of different onscreen keyboard layouts using EEG signals. In: *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 480–486 (2013). <https://doi.org/10.1109/SMC.2013.88>
36. GSR R-Notebooks: Processing in iMotions and algorithms used (Latest Version) (2021). <https://help.imotions.com/hc/en-us/articles/360010312220-GSR-R-Notebooks-Processing-in-iMotions-and-algorithms-used-Latest-Version>. Accessed 6 Jan 2021
37. R Notebooks (EDA): GSR Epoching (2021). <https://help.imotions.com/hc/en-us/articles/360013685940-R-Notebooks-EDA-GSR-Epoching>. Accessed 6 Jan 2021