applied sciences

MDPI

Article

Analysing Virtual Labs Through Integrated Multi-Channel Eye-Tracking Technology: A Proposal for an Explanatory Fit Model

María Consuelo Sáiz-Manzanares [1,*], Raúl Marticorena-Sánchez [2], Javier Sáez-García [3] and Irene González-Díez [4]

[1] Department of Health Sciences, GIR DATAHES, UIC JCYL Nº 348, Universidad de Burgos, 09001 Burgos, Spain
[2] ADMIRABLE Research Group, Department of Computer Engineering, Higher Polytechnic School, Campus Vena, Universidad de Burgos, 09006 Burgos, Spain; rmartico@ubu.es
[3] Higher Polytechnic School, Campus Milanera, Universidad de Burgos, 09001 Burgos, Spain; jsg1013@alu.ubu.es
[4] Department of Health Sciences, GIR DATAHES, Universidad de Burgos, 09001 Burgos, Spain; igdiez@ubu.es
* Correspondence: mcsmanzanares@ubu.es

Abstract: This study deals with an analysis of the cognitive load indicators produced in virtual simulation tasks through supervised and unsupervised machine learning techniques. The objectives were (1) to identify the most important cognitive load indicators through the use of supervised and unsupervised machine learning techniques; (2) to study which type of task presentation was most effective at reducing the task's intrinsic load and increasing its germane load; and (3) to propose an explanatory model and find its fit indicators. We worked with a sample of 48 health sciences and biomedical engineering students from the University of Burgos (Spain). The results indicate that being able to see the task before performing it increases the germane load and decreases the intrinsic load. Similarly, allowing students a choice of presentation channel for the task respects how they process information. In addition, indicators of cognitive load were found to be grouped into components of position, speed, psychogalvanic response, and skin conductance. An explanatory model was proposed and obtained acceptable fit indicators.

Keywords: eye tracking; galvanic skin response; cognitive load; simulation tasks; machine learning techniques

check for updates

Citation: Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Sáez-García, J.; González-Díez, I. Analysing Virtual Labs Through Integrated Multi-Channel Eye-Tracking Technology: A Proposal for an Explanatory Fit Model. *Appl. Sci.* **2024**, *14*, 9831. https://doi.org/10.3390/app14219831

Academic Editors: Vassilios Krassanakis, Erwan David and Olivier Le Meur

Received: 5 September 2024
Revised: 19 October 2024
Accepted: 22 October 2024
Published: 28 October 2024

1. Introduction

This paper will first address the state of the art of the concept of cognitive load during task or problem solving. Secondly, it will address the challenge of interpreting cognitive load indicators through statistical and machine learning algorithms (supervised and unsupervised). Finally, a suggested model of analysis will be analysed as an advance in the knowledge of cognitive load and its application to fieldwork.

1.1. Cognitive Load Definition and Measurement

In recent years, the measurement of cognitive load with respect to task performance or problem solving has become a necessary process, as it helps us understand performance and facilitates an analysis and adjustment of how tasks or problems are presented. Interest arises from a variety of fields and contexts (the educational and instructional sector, marketing, driving, aviation, business, etc.). However, there are many challenges involved in what seems at first to be manageable.

The concept of cognitive load is based on the idea that cognitive resources are limited. Cognitive load develops as a complex interaction between the different demands of tasks and the cognitive processes required to process and execute them, and a subject's performance often decreases when cognitive demands are either very low or very high. The techniques for measuring cognitive load can be classified into four main categories: (1) performance-based, (2) subjective, (3) behavioural and (4) physiological [1–3]. Performance can be measured through learning outcomes, subjectivity through learning history and the use of scales, and behaviours through behavioural analyses. When it comes to physiological components, various indicators can be applied to measure physiological responses during task performance or problem solving. For example, research using eye tracking technology shows that the frequency of fixations, saccades and pupil diameter appear to be suitable indicators for predicting task difficulty at an early stage [4].

The theory of cognitive load is related to studies in the field of developmental educational psychology and instructional cognitive psychology, specifically the study of the relationship between working memory and long-term memory (LTM) [5]. The theory first appeared in 1976 with Sweller's studies [6]. Subsequently, in the late 1980s, the concept of cognitive load was used to explain performance on different tasks with respect to the use of various forms of information processing. A distinction was made between the intrinsic cognitive load (which refers to the complexity of the knowledge to be processed by the learner and is independent of how the task is presented) [7] and extrinsic cognitive load (referring to how the task is presented) [8,9]. Cognitive load seems to be related to divided attention with respect to information received through different channels (visual, auditory, or both). This information must be integrated in order for the learning agent to gain an understanding of a task [9,10]. The comprehension process also seems to be influenced by the way the task is presented [11]. These results are based on Baddeley's theory [12], according to which memory is divided into different types of storage [short-term memory (STM), working memory, and LTM]. For Baddeley [12], working memory consists of an auditory subsystem and a visual subsystem.

Until the second half of the 1990s, research on cognitive load theory focused on instructional designs seeking to reduce the extrinsic cognitive load, since at that time the intrinsic load was difficult to measure or manipulate. Later, the concept of germane cognitive load was used to refer to the mental resources that learners deploy to deal with the intrinsic cognitive load of learning tasks [8]. It was introduced in reference to the development of cognitive schemas in LTM on knowledge integration. Subsequently, studies by Sweller [13] showed that integrating information through different channels (visual or auditory) does not always lead to better learning. The explanation was that eliminating redundant information can be beneficial in the learning process.

Sweller tried to integrate cognitive load theory into the broader theoretical framework of evolutionary theory [13]. This new view was explained by a representation of the cognitive architecture in which there are different memory stores. More specifically, the working memory store has a limited capacity when dealing with new information and cannot combine or contrast more than four pieces of information at a time [13,14]. In addition, information in working memory is lost in approximately 20 s [15]. In LTM, information is organised into higher-order units called cognitive schemas, according to Baddeley [12]. When the number of information units that are processed simultaneously exceeds the capacity of working memory, these units must be combined into cognitive schemas in order to understand the information [16]. In short, each cognitive process that requires conscious control places a cognitive load on working memory. In this framework, germane load helps in the construction and automation of schemas. Therefore, according to the theory of cognitive load, the germane load should be increased as much as possible.

In addition, explanations of cognitive load should consider Vygotsky's theory of the Zone of Proximal Development (ZDP) [17], where the success of learning new information is related to the cognitive links that the learner makes with respect to similar information. The construction of new information will be based on this interaction [13,17]. Likewise, the theory of cognitive load is currently explained by a theory of instruction [18] that differentiates between primary biological information. This refers to information that we as members of the species have about different abilities (speaking, walking, etc.) that we execute unconsciously. However, when the learner presents certain delays in development or is subjected to socio-environmental deprivation factors, there may be problems in the acquisition of these skills. In addition, this theory of cognitive load includes secondary biological information, which requires conscious effort to successfully learn various skills such as reading, writing, arithmetic, etc. Acquiring this knowledge needs the application of various specific metacognitive strategies (orientation, planning, evaluation, and elaboration of information).

Putting all that together, the starting point is therefore an explanatory scheme that begins with an input of information where the learner first of all perceives information (this can be visual, auditory, tactile, or any combination thereof). The learner then selects some of the information and starts processing it. Some of this information passes to STM and working memory (this store cannot remember more than 7 items of novel information and cannot process more than 3–4 items, and these actions cannot last more than about 20 s). Subsequently, the retained information will pass to LTM (this store has more storage capacity than STM and the information remains more stable). LTM would therefore be the key to the individual difference in human processing, as each learner will relate new information to existing information in a different way. The load here is both intrinsic and extrinsic. In order to achieve good learning, learners' motivation has to be enhanced in the way information is presented by seeking novelty and avoiding redundancy. On the other hand, improving the retention of novel information requires facilitating practice and automation of task resolution procedures. Each of these effects will facilitate the acquisition of information stored in LTM by reducing the load on working memory [18], although all of these processes are subject to individual differences in learners [18].

This means that understanding the human cognitive architecture is important for developmental educational psychology and instructional cognitive psychology, as knowledge of the types of processing will provide the teacher with information for tailoring the design of learning tasks [18]. Therefore, knowledge of the cognitive load in the performance of a given task is important for adjusting the task to each learner's performance possibilities. However, this remains a challenge, as the same task may need to be presented in different ways for each type of learner, and learners may experience different cognitive loads for the same task [19].

Research is currently ongoing about the cognitive load experienced in virtual environments that apply simulation tasks [19]. These studies show that cognitive load is a useful indicator for analysing learning processes in simulated environments. They also differentiate between intrinsic cognitive load, task-specific germane cognitive load, and extraneous cognitive load. If the total cognitive load (the sum of the three types) exceeds the capacity of working memory, performance may suffer [20]. Another classification of cognitive load differentiates between primary cognitive load (referring to domain-specific task performance) and secondary cognitive load (referring to metacognitive performance that supports primary performance). If the secondary cognitive load is activated, performance may be better but the total cognitive load is increased, and so the relationship between cognitive load and performance may be positive or negative depending on the type of task, the type of learner, the task design, and the context of task performance [21].

Research about performance analysis in simulated learning environments has shown psychophysiological indicators of cognitive load to be the most reliable. These include pupil dilation (an involuntary response reflecting noradrenergic activity of the locus coeruleus that regulates arousal, cognitive activity, and emotion), which can indicate the cognitive load related to the emotion that the task produces in the learner [22,23]. However, this measure may be affected by factors other than an individual's processing, such as the type of lighting. Other indicators of cognitive load include the duration and total time of fixation [22], the number of saccades, and the amplitude of the saccades [24,25]. These are all static indicators, but dynamic indicators should also be considered. Dynamic indicators refer to the route that each learner takes in the task resolution process. The route refers to the steps followed in order of priority and the duration of each step in the spatial coordinate space during execution. An analysis of each learner's tracking pattern on the multimedia interface used in the simulation environments can give information about how this information has been retained in working memory and how it has been processed in LTM. The type of interface can increase or decrease a task's cognitive load, depending on the individual learner [26].

In summary, physiological measurements during task performance can provide information on different indicators such as pupil diameter [27] or the galvanic skin response (GSR), which measures skin conductance by recording the sweat secreted from the sweat glands [28]. This can be divided into tonic components, giving information on the level of skin conductance, and phasic components, providing the skin conductance response (SCR), a rapidly changing signal that takes one to five seconds to reflect the response to a stimulus [29]. These two measures, GSR and SCR, are considered reliable indicators for measuring cognitive load levels [27,29], although they can be contaminated by other external factors. As noted above, pupillary activity can be subject to illumination-related factors [30]. Similarly, skin conductance indicators may be affected by the user's physical activity (e.g., GSR-equipped hand movements) [29]. Therefore, to prevent external or internal factors from affecting the physiological measurement of cognitive load, these external variables should be controlled as much as possible [31]. Moreover, current lines of research are aimed at using several indicators to measure cognitive load, understanding this as a multimodal approach [31,32]. Such an approach has many advantages, as using various indicators of cognitive load can mitigate the potential limitations of individual measures and thus provide a more comprehensive analysis of the cognitive load experienced by different users. In addition, recording physiological measures allows real-time data collection, facilitating a more objective analysis of cognitive load [31]. Another factor that is relevant to cognitive load studies is the design of the materials (content, interface, etc.) used in the task to be solved [33]. Table 1 presents a list of possible variables that can influence learners' performance of a task or problem solving. An analysis of these variables may provide researchers with information about each learner's characteristics with respect to the intrinsic and germane cognitive loads experienced in each task.

Table 1. List of elements to be considered in the task performance space and measurement indicators.

Elements to Consider	Measurement
Task characteristics, indicators of the extrinsic cognitive load (form of task presentation (visual—image, text, etc.)) and extrinsic cognitive load written, auditory, all), multimedia learning environment simulation, task execution schedule, etc.	Description of the type of task Analysis of the presentation of the task
Physical space [characteristics of the space in which the task takes place (lighting, space, etc.)].	Indicators for brightness, temperature, noise, time, and time of day
Learner characteristics (age, gender, prior knowledge, motivation towards the task, way of responding, way of solving, type of reasoning applied, etc.).	Variables assigned (age, gender, etc.) Analysis of prior knowledge Cognitive style and learning style
Physiological indicators of the intrinsic cognitive load	Pupil diameter Fixations Saccades GSR amplitude SCR count
Perception of the germane load	Performance indicators in the execution of the task

1.2. Machine Learning Algorithms Used in the Interpretation of Cognitive Load

As discussed above, many different factors can influence the performance of a task or the resolution of a problem. Measuring them provides a huge amount of data to be processed. Various statistical and machine learning techniques can be used to carry out this processing. These techniques are divided into supervised techniques using labelled data to train models (within which we can differentiate between classification and regression techniques); unsupervised techniques that do not require labels and are used to discover hidden patterns in the data; and semi-supervised learning (combines the most appropriate supervised and unsupervised approach in each case of study).

The techniques include algorithms such as logistic regression [34]. These can help in the classification of intrinsic cognitive load into levels (low, medium, and high) [19]. A probabilistic study from a percentile analysis can also be used for this purpose [35]. Similarly, supervised machine learning algorithms can be used—where it is advisable to select features beforehand in order to detect the most influential ones [19]. Next, regression and/or classification techniques can be applied. Regression techniques may include linear regression, Support Vector Machine (SVM), Decision tree, and Neural Network algorithms. Classification techniques that can be used include discriminant analysis and *k*-Nearest neighbour (KNN) algorithms, among others.

Unsupervised learning techniques (clustering) include applying algorithms such as *k*-Means, *k*-Medoids, and s-Means [36]. Therefore, before data processing, it is important for noise to be removed and classifiers to be found. The aim is to determine the best indicators of cognitive load in each task [19]. Principal component analysis can also be applied to find a feature matrix [37]. For example, many workload studies have used the SVM algorithm in multiclass classification [36], whereas other studies have used KNN, Naive Bayes, Random Forest, SVM [38–40], and neural network-based models to infer cognitive load levels [41,42]. Classifier effectiveness can be determined using accuracy, recall, and precision indicators in the different algorithms used [41–43]. Fit can be examined using goodness-of-fit index models, which include various fit tests (NFI = normed-fit-index; RMSEA = root-mean-square error of approximation; SRMR = standardised root-mean-square residual; TLI = Tucker–Lewis index; CFI = comparative fit index; AIC = Akaike information criterion; and ECVI = parsimony index) [44].

In contrast, as already indicated, unsupervised learning clustering techniques can be used to learn the clustering profile with respect to different indicators of cognitive load without applying a prior classifier [45]. Algorithms can also be used to perform pattern analysis. These include distance analysis through string edit methods [46], which are applied in the study of patterns detected with a scan path or gaze point. Similarly, more traditional statistical algorithms such as Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), and Multivariate Analysis of Variance (MANOVA) can be used to determine the significant differences in indicators of cognitive load between different types of users [47]. Table 2 presents a summary of the algorithms and their usefulness in the study of cognitive load.

Table 2. Different algorithms and their usefulness for measuring cognitive load.

Contrast Tests	Usefulness in Measuring Cognitive Load
Feature selection [19,37]	The use of feature selection algorithms provides information in a hierarchical order on which features are most relevant.
Logistic regression [34] Probabilistic study (percentile analysis) [36].	Helps in the detection of intrinsic cognitive load levels (low, medium, and high) [19]. Allows a determination of the level of execution (low, medium, and high).
Predictions (linear regression, SVM, Decision tree, Neuronal Networks)	Provide information on cognitive load indicators that may be more relevant in explaining, e.g., learning outcomes.
Classification (analysis, discriminant analysis, KNN, SVM, Naive Bayes, Random Forest [38–40], and Neural networks [41,42])	The classifiers provide information on the most important cognitive load indicators with respect to ranking groups of variables in relation to a reference variable, such as academic performance.

Table 2. *Cont.*

Contrast Tests	Usefulness in Measuring Cognitive Load
Clustering (*k*-Means, *k*-Medoids, s-Means) [45], string-edit Methods [46]	Determines interaction profiles of task performance without prior grouping variables. Identifies route maps in different types of learners.
Adjustment indicators (accuracy, recall, and precision) [41–43], goodness-of-fit indices (NFI, RMSEA, SRMR, TLI, CFI, AIC, ECVI) [44]	Determine the effectiveness, fit and precision when comparing the results obtained with different algorithms.
Parametric statistics (ANOVA, ANCOVA, MANOVA and effect sizes) [47] If the sample does not have a normal distribution, similar tests would be applied in a non-parametric context (Mann–Whitney U test, Kruskal–Wallis, Wilcoxon signed-rank test, etc.)	Determine whether there are significant differences in terms of variables considered as independent variables on variables considered as dependent variables. Also, determine the size of the significant effect.

1.3. Proposed Models for Measuring Cognitive Load

Taking the above into account, the current challenge is to create models to explain cognitive processing in different types of learners during the same task [48]. As noted above, various algorithms have been advanced for measuring the cognitive load of different users during the execution of different tasks [49]. Proposed explanatory and analytical models are based on the construction of multimodal machine learning models that help to estimate people's cognitive loads. However, one potential limitation of these models is that they usually use data extracted from the same scenario. Therefore, the models would need to be tested in learning contexts other than the source application and include information on the execution of different learning tasks. Furthermore, predictive results should be accompanied by estimates of uncertainty (percentage error) in order to reduce misclassification [50].

In summary, the current research focuses on finding robust models that can predict certain situations and facilitate early intervention [51]. These models are based on using machine learning techniques because, as noted above, the assessment of cognitive load is complex. Current research [52] indicates that machine learning techniques such as Deep Learning (DL) are a good resource for finding the most relevant information, allowing a more reliable classification. In particular, hybrid DL architectures that integrate convolution operators and Long Short-Term Memory (LSTM) have been found to produce significantly better results than other models. Similarly, the proposed models should be tested in domains other than the source context, in addition to taking into account individualised differentiation with respect to users [53]. Models that apply inverse reinforcement learning are also proving particularly effective at explaining complex cognitive tasks [54]. Likewise, hyper long-short-term-memory-based modules from time-series multimodal information models have been shown to be effective at explaining the cognitive load in driving tasks [55]. The use of data fusion techniques [56] and Transformer models are likely to be particularly useful resources in explaining the cognitive load experienced by users in the execution of different tasks [57–59]. In recent years, various studies have applied a variety of techniques, such as the Gaussian Mixture Variational-Based Transformer Domain Adaptation Fault Diagnosis method (applying it to diagnosing faults in bearings) [60], and the Prototypical Contrastive-Based Domain Adaptation method to achieve learning-based fault diagnosis under variable working conditions [61].

In summary, modelling is important in different research fields (neurodegenerative diseases, cancer processes, autism spectrum disorders, learning disabilities, etc.) where an early differential diagnosis is essential, as early therapeutic or instructional intervention is critical for a good prognosis. Therefore, cognitive load indicators (biomarkers) need to be identified, and predictive and classification models need to be proposed [62]. Furthermore, the results of current research advocate the use of an approach that includes several indicators of cognitive load collected through different devices. This is because this working methodology increases the reliability of cognitive load detection and the identification of individual differences. Similarly, attention needs to be paid to the design of the task and the development of the interface where it is applied. This area of research faces challenges such as noise reduction with respect to the signal collection devices and the extrapolation of the results produced in laboratories [19].

Based on the state of the art above, the objectives of this study were as follows:

Objective 1. To select characteristics with respect to different indicators of cognitive load related to the performance of simulation tasks in virtual environments.

Objective 2. To determine participants' levels of cognitive load (low, medium, and high).

Objective 3. To apply supervised machine learning prediction techniques to determine how the variables of presentation type 1 (not having previously watched the lab vs. having done so) and presentation type 2 (visual–auditory vs. visual) from the virtual laboratories predict the indicators of cognitive load and learning results.

Objective 4. To apply supervised machine learning techniques for classification in order to determine the most representative classifiers of cognitive load.

Objective 5. To apply unsupervised machine learning techniques to identify possible groupings without a prior assignment variable with respect to cognitive load indicators and learning outcomes.

Objective 6. To determine whether there are significant differences in the indicators of cognitive load and learning results depending on the type of task and the way it is presented (visual–auditory vs. visual) in the participating groups.

Objective 7. To propose an analysis model for the markers of cognitive load.

2. Materials and Methods

2.1. Participants

We worked with a convenience sample of 48 undergraduate students in their final years at the University of Burgos: 21 were studying for a degree in Occupational Therapy, of whom 20 were women and 1 was a man (Group 1); 10 were studying for a degree in Biomedical Engineering (1), of whom 8 were women and 2 were men (Group 2); and 17 were studying for a degree in Biomedical Engineering (2), of whom 9 were women and 8 were men (Group 3). Groups 2 and 3 were different groups, although they shared the same degree and academic level. The mean age of the participants was between 21–23 years old. Participation in the study was voluntary, required written informed consent, and was without financial compensation.

2.2. Instruments

(a) Tobii pro lab (Stockholm, Sweden) version 1.194 with 64 Hz.
It is eye-tracking software that provides a visual user interface and dedicated software functions that support the researcher in all phases of an eye-tracking experiment, from test design and recording to analysis. An image of the device used is shown in Figure 1 below.

(b) Laptop computer. Falkon i7. Screen 15.6"; processor Intel core I7-9750H; hard disk 512 Gb SSD; memory 16 Gb of RAM; Geforce RTX 2070 MAX-Q GDDR6 8 GB; Windows 10 Professional; connections: HDMI/USB-C/USB 3.1/RJ45 × 1/HDMI × 1/3.5 mm Jack/USB-A 3.1 × 4/Thunderbolt × 1—Bluetooth/WiFi/Gigabit Ethernet. Display resolution of 1080 × 1920.

(c) Shimmer3 GSR+ (galvanic skin response) (Europe Research Ltd., Dublin, Ireland) single-channel galvanic skin response data acquisition unit (measures electrodermal resistance (EDR)/electrodermal activity (EDA)). The GSR+ unit is suitable for measuring the electrical characteristics or conductance of skin. Traditional galvanic skin response theory is based on the idea that the electrical resistance of the skin varies with the state of the sweat glands. Sweating in the human body is regulated by the autonomous nervous system (ANS). More specifically, if the sympathetic branch (SNS) of the ANS is excited, the activity of sweat glands increases, which in turn increases skin conductance. This device is compatible with Tobii pro lab version 1.194 and allows for the integration of records. Figure 2 shows the device and how it is placed to record EDA.

(d) Virtual lab for the resolution of clinical cases. This lab was developed within the European project "Specialized and updated training on supporting advance technologies for early childhood education and care professionals and graduates", eEarlyCare-T-No. 2021-1-ES01-KA220-SCH-000032661. In this study, the virtual lab is a representation of a dialogue between two avatars: a student and a teacher.
Two labs were used in this study. In Group 1 and Group 2, a general simulation virtual lab was applied (this referred to content that could be used in both degrees, as they were general and about patient intervention processes). This lab is extensively described in the study by Sáiz-Manzanares et al. [32] (see Figure 1 in that publication). In addition, a specific lab with biomedical engineering content was applied to Group 3. This lab can be found in Supplementary Materials (Figure S1). Both virtual labs can be accessed and used free of charge by logging into the Virtual Classroom in the eEarlyCare-T project https://www2.ubu.es/eearlycare_t/en/project (accessed on 16 August 2024).

(e) Learning Strategies Scales (ACRA) by Román and Gallego [63]. This is an instrument that has been widely tested in research. It identifies 32 strategies at different points in the information processing cycle and has reliability indicators between α = 0.75 and α = 0.90, and content validity indicators between r = 0.85 and r = 0.88. In the present study, we applied the metacognitive skills subscale, which includes 17 strategies on the use of metacognitive skills in problem-solving tasks that are distributed over three subscales: Self-knowledge, Planning, and Self-Evaluation.

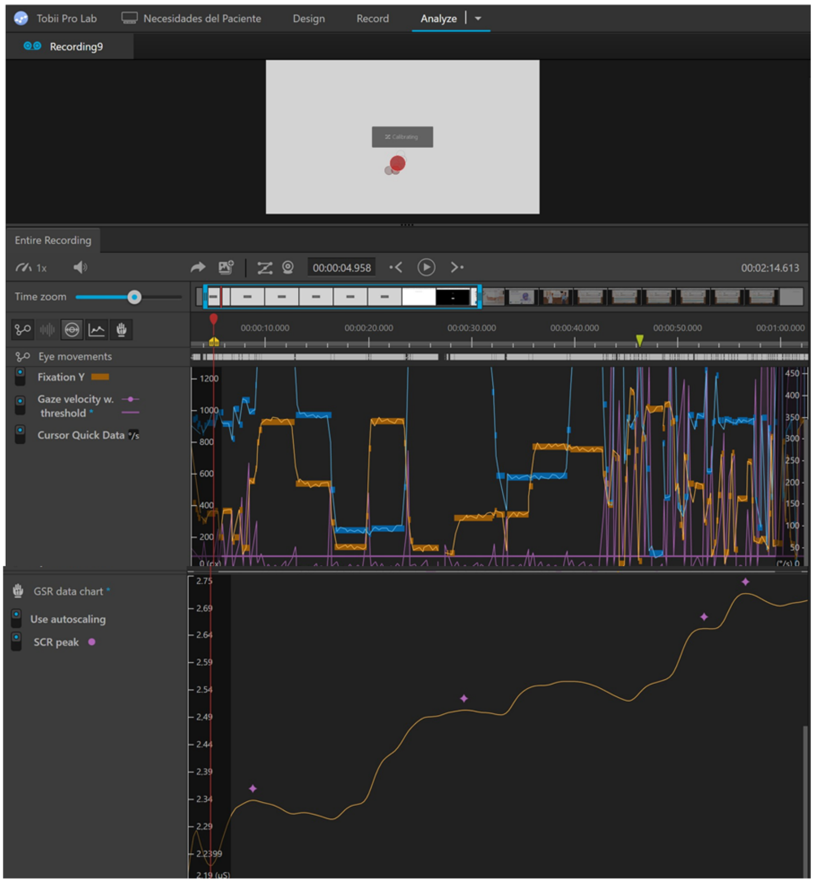

Figure 1. Example of a test performed with Tobii pro lab version 1.194 with 64 Hz, recording eye movements and GSR.

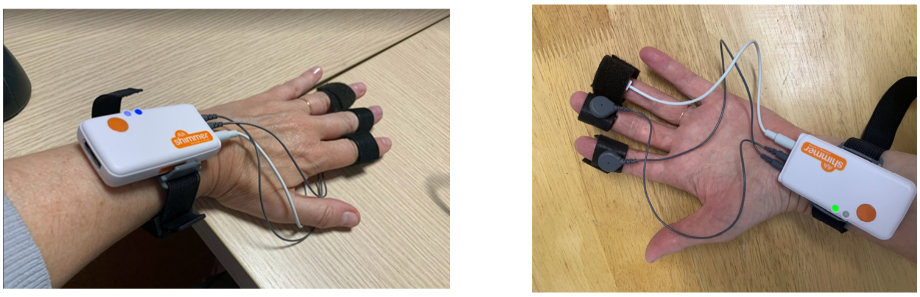

Figure 2. Example of the Shimmer3 GSR+ device.

2.3. Procedure

Prior approval was obtained for the study from the Bioethics Commission of the University of Burgos: No. IO 03/2022 (for carrying out the study) and No. IO 04/2022 (for application of the labs). Written informed consent was then obtained from all participants, their participation was voluntary and without financial compensation. The virtual labs were then implemented for all three groups throughout the second semester. In all cases, lighting and temperature conditions were equivalent (the light and temperature conditions were stable, as the studies were always carried out at the same time with the same light conditions and the temperature was always regulated by a thermostat at 21°). The time each participant spent viewing each lab was discretionary, as they were able to spend as long as they felt necessary to watch the scenes making up each lab. The labs could be watched with audio (an avatar narrating the written text presented in each lab scene) or with only the image of the scene showing the avatars and the text of their dialogue. At the beginning of the lab, each participant was informed that they could voluntarily choose how they wanted to watch (with audio vs. without audio). Group 1 and Group 2 saw the same lab, which included general health content (which can be found in Figure 1 in the paper by Sáiz-Manzanares et al. [32]). The participants in these groups had not previously seen the lab. Group 3 watched a lab that included specific biomedical engineering content (specifically referring to the functionality of using eye-tracking technology and its measurement indicators) (see Figure S1 in Supplementary Materials document). Participants in this group were able to see the lab before watching it again in the in the experimental room. After all three groups had watched the lab in the experimental room, a test was administered to test the knowledge acquired. The test was multiple choice with a single correct answer for each question. The test given to Group 1 and 2 is provided in the Supplementary Materials, Figure S2, and the test given to group 3 is in Figure S3. As the number of questions in the check tests were not equal in order to compare the answers, the answers were categorised into correct versus not correct answers.

2.4. Data Analysis

Table 3 presents the tests and instruments used to test the research objectives.

Table 3. Research objectives, counterfactuals, and analytical tools.

Research Objectives	Test	Test Instrument
Objective 1. To select characteristics with respect to different indicators of cognitive load related to the performance of simulation tasks in virtual environments.	Feature selection	WEKA v.3.8.5 [64]
Objective 2. To determine participants' levels of cognitive load (low, medium, and high).	Percentiles (25, 50, and 75)	SPSS v. 28 [65]
Objective 3. To apply supervised machine learning prediction techniques to determine how the variables of presentation type 1 (not having previously watched the lab vs. having done so) and presentation type 2 (visual–auditory vs. visual) from the virtual laboratories predict the indicators of cognitive load and learning results.	Multiple linear regression	SPSS v. 28 [65]

Table 3. *Cont.*

Research Objectives	Test	Test Instrument
Objective 3 (continued) To apply principal component analysis to determine the groupings of the cognitive load indicators.	Principal component analysis	SPSS v. 28 [65]
Objective 4. To apply supervised machine learning techniques for classification in order to determine the most representative classifiers of cognitive load.	Discriminant analysis KNN	SPSS v. 28 [65]
Objective 5. To apply unsupervised machine learning techniques to find possible groupings without a prior assignment variable with respect to cognitive load indicators and learning outcomes.	*k*-means	SPSS v. 28 [65]
Objective 6. To determine whether there are significant differences in the indicators of cognitive load and learning outcomes depending on the type of task and the way it is presented (visual–auditory vs. visual) in the participating groups.	Two-factor fixed-effect MANOVA	SPSS v. 28 [65]
Objective 7. To propose an analysis model for the indicators of cognitive load.	Structural equation modelling	AMOS v. 29 [66]

3. Results

3.1. Objective 1: To Select Characteristics with Respect to Different Indicators of Cognitive Load Related to the Performance of Simulation Tasks in Virtual Environments

WEKA algorithm [58] Evaluator, weka.attributeSelection.CfsSubsetEval, evaluates the value of a subset of attributes taking into account the individual predictive ability of each feature together with the degree of redundancy between features. More information is available at https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CfsSubsetEval.html (accessed on 20 August 2024). The search is weka.attributeSelection.BestFirst; this searches the attribute subset space using the greedy hill climbing method augmented with a backtracking function. More information is available at https://weka.sourceforge.io/doc.dev/weka/attributeSelection/BestFirst.html (accessed on 20 August 2024). Forty-eight instances and 10 attributes were analysed (mean duration, fixation point X, fixation point Y, average pupil diameter, saccade direction, average velocity, peak velocity, saccade amplitude, average GSR, and SCR count). The results were the merit of best subset found, 0.75, and the features selected were the fixation point X and average GSR.

3.2. Objective 2: To Determine Cognitive Load Levels (Low, Medium, and High) in the Participating Groups

Descriptive statistics, maximum and minimum values, and percentiles were calculated for each of the characteristics. The 25th percentile indicates a low value, the 50th percentile an intermediate value, and the 75th percentile a high value, see Table 4.

Table 4. Descriptive statistics and percentiles for cognitive load indicators.

Indicators of Cognitive Load	Mean	SD	Minimum	Maximum	Percentiles		
					25	50	75
Group 1							
Mean duration	140.72	34.77	105.33	224.75	118.63	140.72	174.36
Fixation point X	994.84	288.42	750.66	1924.86	798.22	994.84	1227.70
Fixation point Y	866.17	195.53	635.68	1247.70	693.81	866.17	1048.09
Average pupil diameter	5311.03	1286.28	3715.67	7991.97	4338.36	5311.03	6558.44
Saccade direction	278,133.97	70,405.30	130,426.22	431,082.91	238,321.58	278,133.97	329,400.50
Average velocity	233,191.56	250,444.66	152,868.39	1,260,057.37	187,812.06	233,191.56	339,436.29
Peak velocity	336,697.42	260,723.55	230,621.84	1,398,529.00	261,396.24	336,697.42	493,273.13
Saccade amplitude	6173.42	2359.82	2020.85	10,407.48	5115.04	6173.42	8437.59
Average GSR	1.47	1.38	0.01	7.44	1.47	1.47	1.47
SCR Count	10.18	5.76	0.00	34.00	10.13	10.18	10.24

Table 4. *Cont.*

Indicators of Cognitive Load	Mean	SD	Minimum	Maximum	Percentiles		
					25	50	75
Group 2							
Mean duration	153.58	35.61	93.25	237.41	136.26	153.58	173.51
Fixation point X	1119.89	527.97	158.17	2424.80	988.74	1119.89	1276.17
Fixation point Y	951.40	438.04	162.18	2023.44	849.83	951.40	1132.69
Average pupil diameter	6371.50	4451.51	1132.41	19,140.53	4432.29	6371.50	7298.63
Saccade direction	319,974.78	115,718.13	41,797.90	397,630.64	267,580.33	319,974.78	375,132.14
Average velocity	229,531.47	94,148.06	49,784.31	382,706.45	199,910.41	229,531.47	266,353.23
Peak velocity	331,430.91	132,544.18	64,173.60	537,397.30	272,005.44	331,430.91	398,652.17
Saccade amplitude	7372.37	2653.55	989.92	9646.79	5714.81	7372.37	8285.77
Average GSR	1.10	1.12	0.01	3.17	0.01	1.10	1.99
SCR count	8.00	11.33	0.00	38.00	0.00	8.00	16.00
Group 3							
Mean duration	250.43	54.94	134.61	399.59	239.68	250.43	286.75
Fixation point X	2919.06	1260.15	1162.95	6185.12	2068.23	2919.06	4091.90
Fixation point Y	2643.52	1307.08	959.91	6157.44	2225.84	2643.52	3557.76
Average pupil diameter	25,055.68	7498.00	11,358.53	40,122.79	19,702.44	25,055.68	29,343.05
Saccade direction	105,956.02	24,936.85	51,991.70	149,285.16	89,490.71	105,956.02	120,311.50
Average velocity	84,465.25	18,890.28	51,017.96	121,059.84	71,890.70	84,465.25	97,068.89
Peak velocity	105,569.23	24,746.30	64,997.00	152,861.24	95,035.23	105,569.23	127,999.83
Saccade amplitude	2006.29	521.67	1300.68	2952.87	1785.50	2006.29	2645.64
Average GSR	0.86	4.11	0.01	14.11	0.18	0.86	4.75
SCR count	11.00	12.14	0.00	39.00	6.00	11.00	24.50

Note: SD = standard deviation; GSR = galvanic skin response, SCR = skin conductance responses.

3.3. Objective 3: To Apply Supervised Machine Learning Prediction Techniques to Determine How the Variables of Presentation Type 1 (Not Having Previously Watched the Lab vs. Having Done So) and Presentation Type 2 (Visual–Auditory vs. Visual) from the Virtual Laboratories Predict the Indicators of Cognitive Load and Learning Results

The presentation type 1 variable (not having previously watched the lab vs. having done so) predicted 76.3% (R^2 = 0.763) of the variance with respect to the cognitive load indicators (see Supplementary Materials, Table S1 for the results of the linear regression analysis). The values of tolerance—the proportion of variance not explained by the rest of the variables (1-R^2), where the higher the Tolerance, the more independent the variable in question—were all less than 1. The variance inflation (VIF) values—which quantifies the intensity of multicollinearity in a normal least squares regression analysis, and the higher the variance of the estimated regression coefficient, the higher the collinearity—were high for the cognitive load indicators average pupil diameter, saccade direction and average velocity.

The presentation type 2 variable (visual–auditory vs. visual) predicted 25.5% (R^2 = 0.255) of the variance with respect to the cognitive load indicators (see Supplementary Materials, Table S2 for the results of the linear regression analysis). In this case, very low values of tolerance were found for the cognitive load indicators fixation point Y, average pupil diameter, saccade direction and average velocity, while there were very high VIF values for the indicators average pupil diameter, saccade direction and peak velocity.

Based on these results, which point to collinearity between the variables, a principal component analysis was carried out to group the cognitive load levels into the most relevant load components (see Supplementary Materials, Figure S4). Three principal components were found (see Supplementary Materials, Table S3 for the matrix of principal components). Component 1 comprised the cognitive load indicators mean duration, fixation point X, fixation point Y, average pupil diameter and saccade direction; Component 2 comprised the indicators average velocity and peak velocity; while Component 3 comprised the indicators average GSR and SCR count.

3.4. Objective 4: To Apply Supervised Machine Learning Classification Techniques to Identify the Most Representative Classifiers of Cognitive Load

A discriminant analysis was carried out to determine the behaviours of the groups with respect to the cognitive load indicators. As Figure 3 shows, the behaviour in cognitive load for Group 1 and Group 2, where the students had not previously watched the lab (which contained general health content), was quite similar to each other, but differed from the behaviour of Group 3, where the students had previously viewed the lab (which was about content specific to the degree). Similarly, all the values of Wilks' lambda (a method for selecting variables by discriminant analysis steps that selects the variables to be added into the equation based on how much they contribute to decreasing Wilks' lambda—at each step, the variable that minimises the overall Wilks' Lambda is added) for the cognitive load indicators was significant except for the average GSR and SCR count indicators (see Table 5).

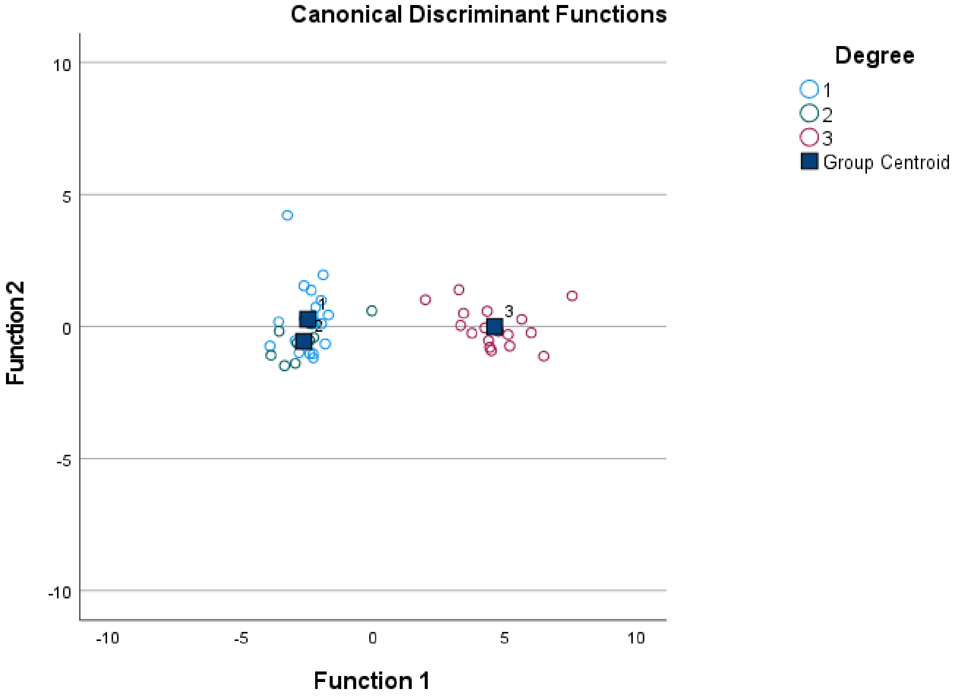

Figure 3. Canonical discriminant functions with respect to the group type variable.

Table 5. Test of equality of group means.

Cognitive Load Indicators	Lambda Wilks	F	df1	df2	pLambda
Mean duration	0.36	40.10	2	45	<0.001 *
Fixation point X	0.38	36.12	2	45	<0.001 *
Fixation point Y	0.39	35.28	2	45	<0.001 *
Average pupil diameter	0.20	92.56	2	45	<0.001 *
Saccade direction	0.35	41.44	2	45	<0.001 *
Average velocity	0.71	9.04	2	45	<0.001 *
Peak velocity	0.62	13.91	2	45	<0.001 *
Saccade amplitude	0.42	31.26	2	45	<0.001 *
Average GSR	0.93	1.79	2	45	0.18
SCR count	0.94	1.43	2	45	0.25

* $p < 0.05$. Note: df = degree of freedom; GSR = galvanic skin response, SCR = skin conductance response.

The KNN algorithm was also applied, giving the most representative cognitive load indicators as fixation point X, fixation point Y, and mean duration.

Figure 4 shows the representation of the focal points with respect to the fixation point X classifier and the distribution of subjects with respect to the other indicators of cognitive load. Figure 5 shows the representation with respect to the fixation point Y classifier and Figure 6 with respect to the mean duration classifier. The numbers in the focal graphs indicate the participant. Thus, it is possible to know the position of each participant with respect to the classifiers and the focal points. This fact is important for being able to design specific tasks for each of them.

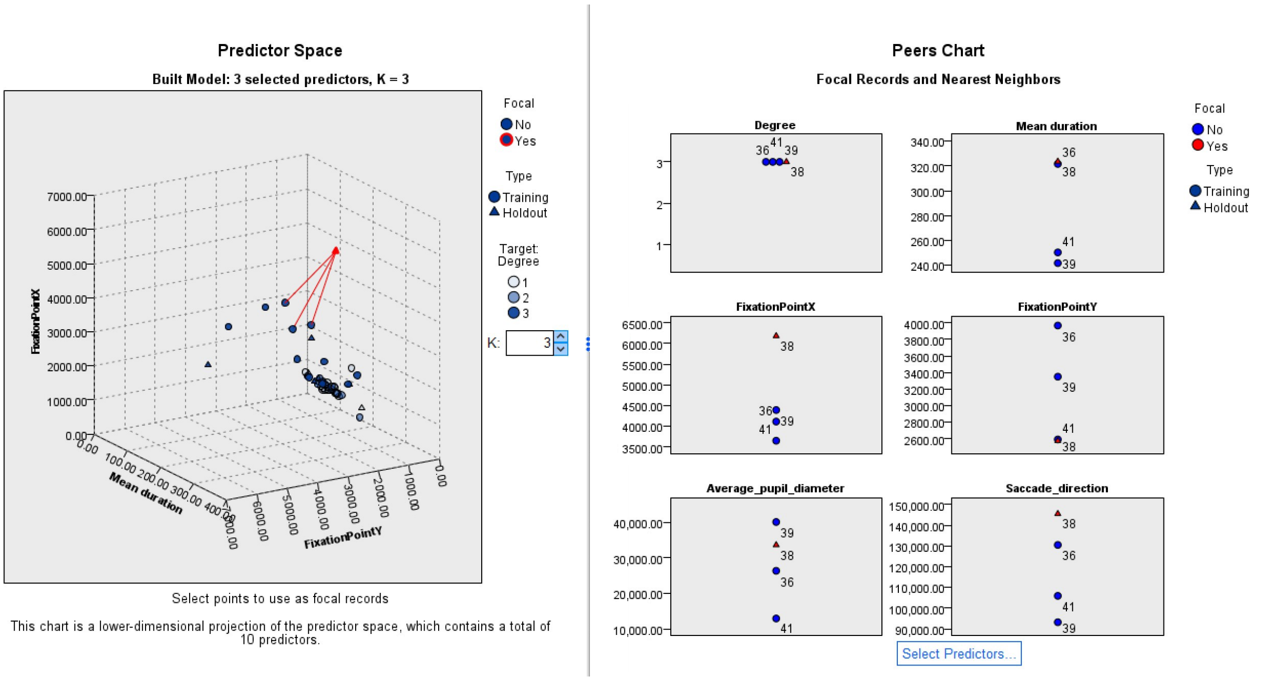

Figure 4. Analysis of the focal points with respect to the fixation point X classifier.

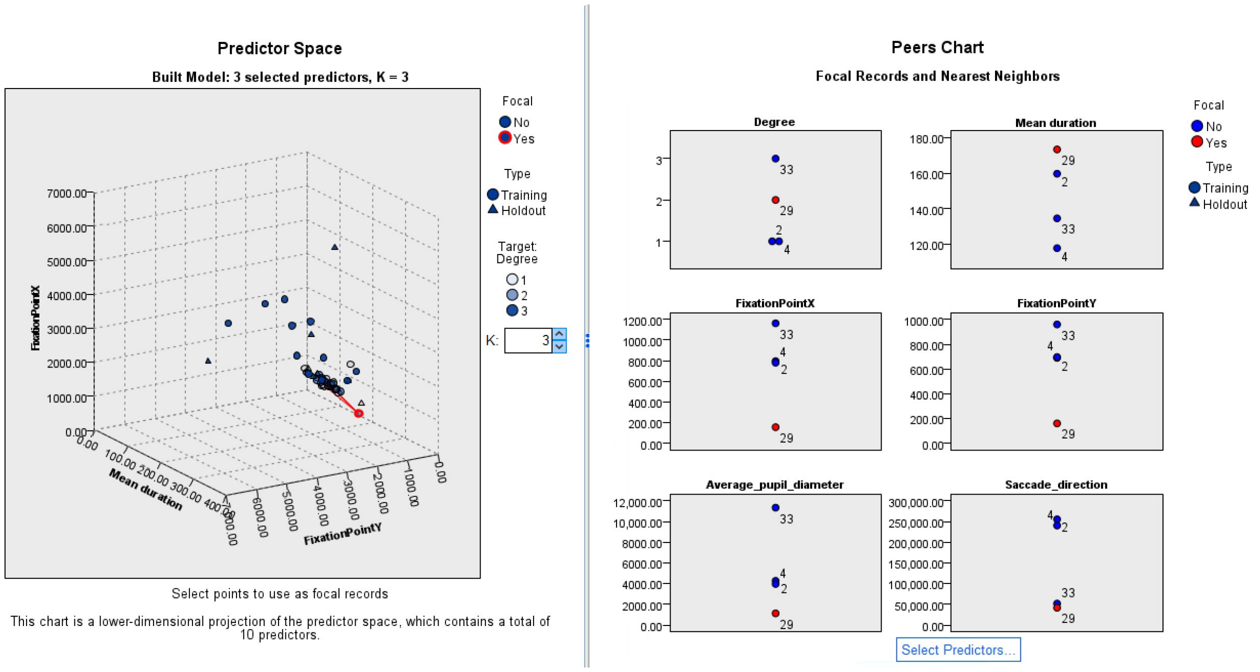

Figure 5. Analysis of the focal points with respect to the fixation point Y classifier.

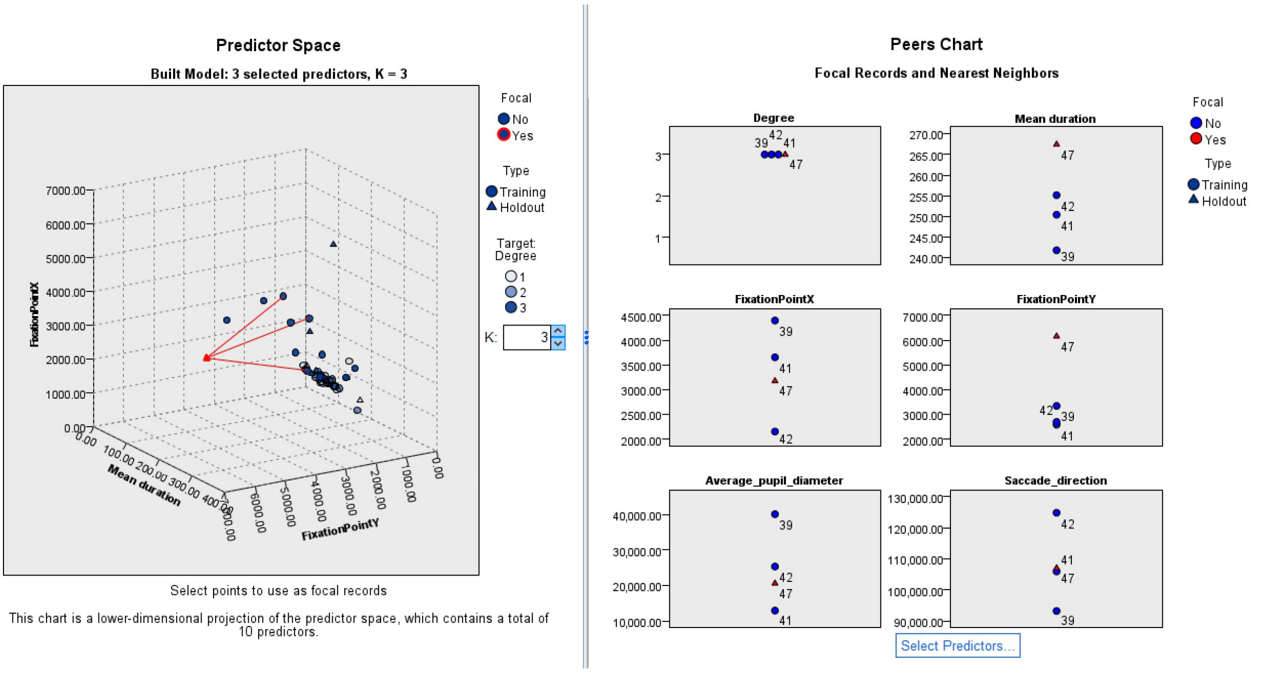

Figure 6. Analysis of the focal points with respect to the mean duration classifier.

3.5. Objective 5: Apply Unsupervised Machine Learning Techniques to Determine Possible Groupings Without a Prior Assignment Variable with Respect to Cognitive Load Indicators and Learning Outcomes

Three clusters were found for the indicators of cognitive load and for the results in the learning check test (see Table 6). ANOVA was also performed to test whether there were significant differences between the clusters in the cognitive load indicators and in the test results (see Table 7). Significant differences were found in all cognitive load indicators except average GSR and SCR count. No significant differences were found in the outcome test.

Table 6. Final cluster centres of *k*-means when *k* = 3 is used.

Indicators of Cognitive Load	Cluster		
	1 n = 22	2 n = 25	3 n = 1
Mean duration	150.58	225.34	172.44
Fixation point X	1165.44	2315.30	994.84
Fixation point Y	997.23	2190.93	921.50
Average pupil diameter	6178.29	17,931.22	6156.46
Saccade direction	322,067.82	140,859.37	130,426.22
Average velocity	298,529.77	109,142.61	1,260,057.37
Peak velocity	406,223.39	147,998.56	1,398,529.00
Saccade amplitude	7562.57	3058.55	2861.72
Average GSR	1.64	2.31	1.47
SCR count	11.54	12.86	10.15
Results test	1	1	1

Note: GSR = Galvanic Skin Response, SCR = Skin Conductance Responses.

Table 7. Cluster ANOVA on indicators (cognitive load and test results).

	Cluster		Error		F	*p*
	Mean Square	df	Mean Square	df		
Mean duration	32,856.22	2	3367.96	45	9.756	<0.001 *
Fixation point X	8,035,738.38	2	1,267,468.79	45	6.340	0.004 *
Fixation point Y	8,584,531.38	2	1,266,594.11	45	6.778	0.003 *
Average pupil diameter	827,482,721.38	2	72,932,182.36	45	11.346	<0.001 *

Table 7. *Cont.*

	Cluster		Error		F	*p*
	Mean Square	**df**	**Mean Square**	**df**		
Saccade direction	196,570,633,958.60	2	4,242,830,417.24	45	46.330	<0.001 *
Average velocity	762,312,710,583.10	2	7,488,341,771.28	45	101.800	<0.001 *
Peak velocity	1,014,921,268,496.76	2	8,945,463,462.81	45	113.457	<0.001 *
Saccade amplitude	121,297,203.11	2	3,563,210.06	45	34.042	<0.001 *
Average GSR	2.76	2	7.55	45	0.365	0.696
SCR count	12.45	2	100.40	45	0.124	0.884
Results test	0.44	2	1.16	45	0.378	0.687

* Note: df = degree of freedom; GSR = galvanic skin response, SCR = skin conductance responses.

Cluster 2 exhibits a longer mean duration compared to the virtual lab display. The mean pupil diameter is also larger, the mean velocity is lower, and the saccade amplitude and maximum velocity have intermediate values.

Cross-tabulation was then performed to test the relationship between cluster membership and group of origin (see Table 8). A contingency coefficient of C = 0.60 $p < 0.001$ was found. All members of Group 3 are in Cluster 2. Members of Group 1 and 2 are distributed between Clusters 1 and 2.

Table 8. Cluster number of cases and degree type cross-tabulation.

Cluster Number of Cases		Degrees			Total
		1	**2**	**3**	
	1	13	9	0	22
	2	7	1	17	25
	3	1	0	0	1
Total		21	10	17	48

3.6. Objective 6: To Determine Whether There Are Significant Differences in the Indicators of Cognitive Load and Test Results Depending on the Type of Task and the Way the Task Is Presented (Visual–Auditory vs. Visual) in the Participation Groups

First, the normality of the sample was examined with respect to the scores on the ACRA metacognitive strategies scale, a test that was applied before the intervention phase in the three groups. Because the sample size was less than 50 (n = 48), the Shapiro–Wilk test was applied, which, as Table 9 shows, indicated a normal distribution. Therefore, parametric statistics were applied to test Objective 6, which in this case was a two-factor multivariate analysis (MANOVA) with fixed effects and eta-squared effect size (η^2) (presentation type 1 (not having previously watched the lab vs. having done so) and presentation type 2 (visual–auditory vs. visual) (see Table 10). Significant differences from for presentation type 1 (prior watching vs. not) were found for all cognitive load indicators except for the average GSR and SCR count. The effect value was small for the cognitive load indicators average velocity, fixation point X, fixation point Y, and peak velocity, intermediate for the mean duration, saccade direction and saccade amplitude, and high for the average pupil diameter. No differences were found in the outcome test. No significant differences were found for presentation type 2 (visual–auditory vs. visual) or for the interaction of the two factors (presentation type 1 and presentation type 2). Descriptive statistics can be found in the Supplementary Materials in Table S4.

Table 9. Shapiro–Wilk normality test.

Metacognitive Strategies	Groups	Shapiro–Wilk test		
		Statistic	df	*p*
Self-knowledge	1	0.96	21	0.53
	2	0.90	10	0.23
	3	0.91	17	0.10
Planning	1	0.92	21	0.08
	2	0.86	10	0.08
	3	0.89	17	0.04
Self-Evaluation	1	0.91	21	0.05
	2	0.97	10	0.86
	3	0.791	17	0.09

$p < 0.05$. Note: df = degree of freedom.

Table 10. Multivariate two-factor fixed effects analysis.

Source	Indicators	F	df	*p*	η^2
Display 1 (no preview vs. display)	Mean duration	25.17	2	0.00 *	0.55
	Fixation point X	17.25	2	0.00 *	0.45
	Fixation point Y	16.90	2	0.00 *	0.45
	Average pupil diameter	54.18	2	0.00 *	0.72
	Saccade direction	30.31	2	0.00 *	0.59
	Average velocity	6.79	2	0.00 *	0.24
	Peak velocity	10.48	2	0.00 *	0.33
	Saccade amplitude	21.68	2	0.00 *	0.51
	Average GSR	0.31	2	0.74	0.01
	SCR count	0.32	2	0.73	0.02
	Results test	0.21	2	0.81	0.01
Visualisation 2 (visual–auditory vs. visual)	Mean duration	0.37	1	0.55	0.01
	Fixation point X	2.21	1	0.15	0.05
	Fixation point Y	2.26	1	0.14	0.05
	Average pupil diameter	0.40	1	0.53	0.01
	Saccade direction	2.27	1	0.14	0.05
	Average velocity	1.35	1	0.25	0.03
	Peak velocity	1.07	1	0.31	0.02
	Saccade amplitude	1.30	1	0.26	0.03
	Average GSR	1.62	1	0.21	0.04
	SCR count	1.54	1	0.22	0.04
	Results test	0.08	1	0.77	0.00
Interaction Visualisation 1 * Visualisation 2	Mean duration	0.90	2	0.41	0.04
	Fixation point X	2.97	2	0.06	0.12
	Fixation point Y	2.21	2	0.12	0.10
	Average pupil diameter	0.52	2	0.60	0.02
	Saccade direction	0.70	2	0.50	0.03
	Average velocity	1.79	2	0.18	0.08
	Peak velocity	2.01	2	0.15	0.09
	Saccade amplitude	0.58	2	0.56	0.03
	Average GSR	0.57	2	0.57	0.03
	SCR count	0.25	2	0.78	0.01
	Results test	0.10	2	0.90	0.00

* $p < 0.05$. Note. GSR = galvanic skin response, SCR = skin conductance responses; df = degree of freedom; η^2 = eta-squared effect size, where values were interpreted according to Cohen [67]: below 0.20 is no effect, between 0.21 and 0.49 is a small effect, between 0.50 and 0.69 is a moderate effect, and ≥0.7 is a large effect.

3.7. Objective 7: To Propose a Model of Analysis of the Markers of Cognitive Load

The results found in testing Objectives 1–6 are summarised in Table 11.

We propose a model for analysing cognitive load indicators that considers the following components. Component 1 includes the cognitive load indicators of position, durations and amplitudes of fixations and saccades, and pupil diameter; Component 2 includes cognitive load indicators of velocity; Component 3 includes indicators of psychogalvanic response and skin conductance; and Component 4 includes the pre-task presentation characteristics and single or multi-channel presentation. A representation of the model is shown in Figure S5. A goodness-of-fit index analysis was then performed to check the fit of the model. The proposed model obtained a better fit in the different fit indices than the stand-alone model (see Table 12). However, the fit was not optimal in some indices, such as RMSEA. Figure 7 shows a representation of the model.

Table 11. Results with respect to the research objectives.

Research Objectives	Results
Objective 1. To select characteristics with respect to different indicators of cognitive load related to performance of simulation tasks in virtual environments.	Feature selection Fixation point X, average GSR
Objective 2. To determine participants' levels of cognitive load (low, medium, and high).	Low $P_{25–49}$ Medium $P_{50–74}$ High $P_{75–99}$
Objective 3. To apply supervised machine learning prediction techniques to determine how the variables of presentation type 1 (not having previously watched the lab vs. having done so) and presentation type 2 (visual–auditory vs. visual) from the virtual laboratories predict the indicators of cognitive load and learning results.	Presentation type 1 [form of task presentation (generalist vs. specific content) and preview vs. no preview] explains 76.3% of the results for the cognitive load indicators. Presentation type 2 [form of task display (visual–audio vs. audio)] explains 25.5% of the results for the cognitive load indicators. However, some of the cognitive load indicators may accumulate redundant information. Three main components were found. Component 1 was related to indicators of duration, position, pupil diameter, saccade direction, and saccade amplitude. Component 2 is related to indicators of velocity, and Component 3 is related to indicators of conductance.
Objective 4. To apply supervised machine learning techniques for classification in order to determine the most representative classifiers of cognitive load.	Differences were found for all indicators of cognitive load depending on the way the task was presented, except for those related to skin conductance. Likewise, the classifiers found were related to the gaze position at point X and Y, and the average duration of cognitive load.
Objective 5. To apply unsupervised machine learning techniques to find possible groupings without a prior assignment variable with respect to cognitive load indicators and learning outcomes.	The clusters found showed greater homogeneity in the group that was able to watch the information before the test. Differences were found in all indicators of cognitive load, except those related to skin conductance and in the results of the knowledge test.
Objective 6. To determine whether there are significant differences in the indicators of cognitive load and learning outcomes depending on the type of task and the way it is presented (visual–auditory vs. visual) in the participating groups.	Significant differences were found with respect to the way the task was presented in all cognitive load indicators, except those related to skin conductance; no differences were found in the results of the knowledge test. No significant differences were found for the display type variable.

Table 12. Standardised estimates of the default model.

Goodness-of-Fit Index	Default Model 1	Independence Model 1	Accepted Value
χ^2	85.98, df = 38 p = 0.00 *.	-	$p > 0.05$ $\alpha = 0.95$
RAMSEA	0.16	0.45	>0.05–0.08
RAMSEA intervals (LO 90)	0.12	0.42	>0.05–0.08
SRMR	0.06	-	>0.05–0.08
NFI	0.85	0.00	0.85–0.90<
TLI	0.90	0.00	0.85–0.90<
IFC	0.90	0.00	0.95–0.97<
AIC	163.98	623.68	The lowest value
ECVI	3.49	13.27	The lowest value
ECVI interval (90%)	2.99	11.70	The lowest value

* $p < 0.05$. Note. df = degree of freedom; χ^2 = Chi-squared; RMSEA = root-mean-square error of approximation; SRMR = standardised root-mean-square Residual; NFI = normed fit index; TLI = Tucker–Lewis index; CFI = comparative fit index; AIC = Akaike information criterion; ECVI = parsimony index.

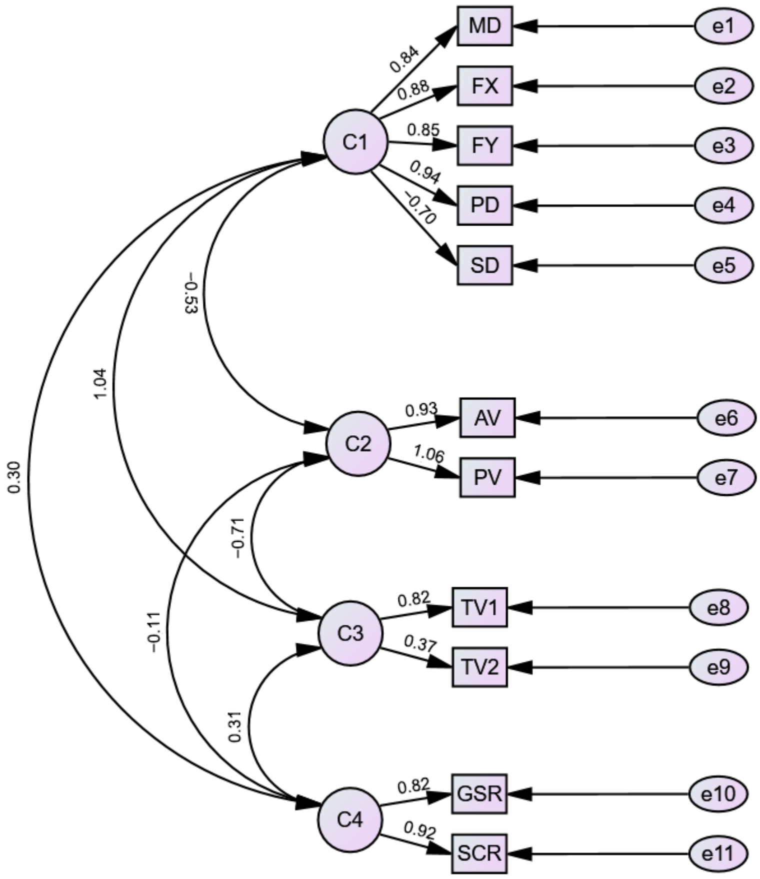

Figure 7. Standardised estimates. Note: Component 1 (C1) includes the following indicators of cognitive load: MD = Mean duration, FX = Fixation point X, FY = Fixation point Y, PD = Average pupil diameter, and SD = Saccade direction. Component 2 (C2) includes the indicators AV = Average velocity and PV = Peak velocity; Component 3 (C3) includes AV = preview, and PV = type of presentation (audio and text vs. audio); and Component 4 (C4) includes the Average GSR and SCR count.

4. Discussion

This study found that the intrinsic cognitive load experienced by students seems to be related to whether they are able to see the task beforehand and to the use of specific learning tasks. However, these conditions did not translate into differences in learning outcomes. Furthermore, these conditions also produced differences with respect to the germane cognitive load, which refers to the construction and automatisation of cognitive schemas in LTM [8,13,17]. Seeing the task beforehand and task-specific content may be related to how the information is processed and the learner's engagement with the task [18,19,21]. One possible explanation is that facilitating knowledge of the information enhances placement of the information in the learner's Proximate Development Zone (PDZ), which helps to improve information processing [17]. At this point, it is assumed that each learner has their own way of processing information [18]; however, by allowing them to see the task before attempting it in the in the simulation lab, it seems that the way information was processed was somehow equalised, with this fact having more explanatory weight than the intrinsic cognitive load that the student might have experienced [20]. The explanation may be related to the fact that the learners' being able to see the information beforehand may help to eliminate redundant information [20] which aids in processing information in working memory and long-term memory [8,9,13,17]. However, this hypothesis will need to be tested in future studies with different tasks and in different groups.

We also found that whether information was presented through one or two channels had an effect on the indicators of cognitive load. This result may be related to the possibility given to students in all groups to voluntarily choose how the task was presented, which may boost the personalisation of the learning process and increase student motivation [18,19,21] as it respects individuality in the way information is processed.

On the other hand, it seems that some of the indicators of cognitive load considered in this study were measuring the same thing with respect to information processing. We found three main components in this study: one related to the cognitive load indicators of duration, position, pupil diameter, saccade direction and saccade amplitude; another related to velocity; and a third related to the psychogalvanic response and skin conductance, i.e., a more emotional component. This result opens the way for future research aimed at studying indicators of cognitive load in terms of explanation from the perceptual and neurological functioning of information processing in different tasks.

Finally, we proposed an explanatory model of cognitive functioning in the execution of simulation tasks. This is intended to highlight the need for applied research to test the generalisability of the research results. More specifically, in this study, we used a data fusion model [56] to analyse the different indicators (cognitive load and type of presentation) and the fit of the results was checked using a goodness-of-fit index. However, future studies will be able to expand on the number of cognitive load indicators, including electroencephalographic recordings; the characteristics of the tasks; and the application of other more powerful models for data processing that include generative artificial intelligence, such as the Transformer model [57–59].

5. Conclusions

Firstly, the results from this study should be considered with caution, as we worked with samples from a single university selected through convenience sampling. In addition, the students voluntarily agreed to participate without receiving any financial compensation, which is a bias in itself. However, the analysis of human processing in learning environments is complex and does not allow for brief tests with very large samples. Therefore, the way to increase generalisability is to replicate research in other contexts with other participants, which will undoubtedly improve the generalisability of the results.

Secondly, bearing in mind that caution is needed when interpreting the results, the study does raise a number of unknowns for future research in this field. It seems that being able to preview tasks in virtual simulation contexts is an important element for homogenising the intrinsic load and increasing the germane load. One possible explanation is that this may help the learner to eliminate the redundant load and improve the processing of information in working memory and long-term memory, which may help increase learning engagement and thus motivation. Another key aspect is to facilitate the learner's choice of information display channel(s). This respects each learner's type of processing (auditory, visual, or both). Therefore, the design of the tasks and the application of the tasks seem to be important elements in improving information processing.

Finally, proposing explanatory models is important. These models must be analysed in terms of their fit and reliability indicators. This is undoubtedly a challenge for developmental educational psychology and cognitive instructional psychology. At this point, the use of generative artificial intelligence similar to what is used in the Transformer models will undoubtedly be of great help in this field of research. Therefore, we stress the importance of working with interdisciplinary teams that include professionals from educational and cognitive psychology, computer science, and information technology in order to effectively address these challenges.

In summary, using characteristic selection techniques can offer researchers important information that helps them understand which of the variables (features in data mining terminology) may be more significant in the study they are conducting. Moreover, an analysis of percentiles with regard to the metrics recorded by multi-channel eye-tracking devices can help researchers to categorise the different types of users (or instances in data mining parlance). Applying supervised machine learning regression techniques gives information about percentage predictions of numerical values, in this case values related to the metrics from multi-channel eye-tracking and GSR devices. Along similar lines, supervised machine learning classification techniques give researchers information about the prediction of previously defined categories. In contrast, non-supervised machine learning clustering techniques help identify hidden patterns, as they do not apply prior labels—unlike supervised learning. Using statistical techniques such as MANOVA allows researchers to determine whether there are significant differences in defined independent variables between the groups being tested. All of these results allow researchers to interpret and develop explanatory models, in this case about information processing during a learning task in a self-regulated virtual lab. Finally, using goodness-of-fit indices gives researchers an idea of how well the proposed model fits.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app14219831/s1.

Author Contributions: Conceptualisation, all authors; methodology, M.C.S.-M., R.M.-S. and J.S.-G.; software, all authors; validation, all authors; formal analysis, M.C.S.-M., R.M.-S. and J.S.-G.; investigation, M.C.S.-M., J.S.-G. and I.G.-D.; resources, M.C.S.-M.; data curation, M.C.S.-M.; writing—original draft preparation, M.C.S.-M.; writing—review and editing, all authors; visualisation, M.C.S.-M.; supervision, all authors; project administration, M.C.S.-M. and R.M.-S.; funding acquisition, M.C.S.-M. and R.M.-S. All authors have read and agreed to the published version of the manuscript.

Funding: The project "Voice assistants and artificial intelligence in Moodle: a path towards a smart university", SmartLearnUni, Call 2020 R&D&I Projects-RTI Type B, MINISTRY OF SCIENCE AND INNOVATION AND UNIVERSITIES, STATE RESEARCH AGENCY, Government of Spain, grant number PID2020-117111RB-I00". Specifically, in the part concerning the application of multichannel eye-tracking technology with university students and the project "Specialized and updated training on supporting advance technologies for early childhood education and care professionals and graduates" (eEarlyCare-T), grant number 2021-1-ES01-KA220-SCH-000032661 funded by the EUROPEAN COMMISSION. In particular, the funding has enabled the development of the e-learning classroom and educational materials.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of University of Burgos (protocol code IO 03/2022 and date of approval 7 April 2022 for the SmartLearnUni project and protocol code IO 04/2022 and date of approval 7 April 2022 for the eEarlyCare-T project).

Informed Consent Statement: Written informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the confidentiality of the data and for ethical reasons and to check the use of the data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brünken, R.; Seufert, T.; Paas, F. Measuring Cognitive Load. In *Cognitive Load Theory*; Plass, J.L., Moreno, R., Brünken, R., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 181–202. [CrossRef]
2. Eggemeier, F.T.; Wilson, G.F.; Kramer, A.F.; Damos, D.L. Workload assessment in multi-task environments. In *Multiple-Task Performance*; Damos, D.L., Ed.; Taylor & Francis: London, UK; Washington, DC, USA, 1991; pp. 207–216. [CrossRef]
3. Johannsen, G. Workload and Workload Measurement. In *Mental Workload*; NATO Conference Series; Moray, N., Ed.; Springer: Boston, MA, USA, 1979; Volume 8, pp. 3–11. [CrossRef]
4. Sevcenko, N.; Appel, T.; Ninaus, M.; Moeller, K.; Gerjets, P. Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: An eye-tracking study. *J. Multimodal User Interfaces* **2023**, *17*, 1–19. [CrossRef]
5. Kirschner, P.A.; Sweller, J.; Kirschner, F.; Zambrano, R.J. From Cognitive Load Theory to Collaborative Cognitive Load Theory. *Intern. J. Comput.-Support. Collab. Learn.* **2018**, *13*, 213–233. [CrossRef]
6. Sweller, J. The effect of task complexity and sequence on rule learning and problem solving. *Br. J. Psychol.* **1976**, *67*, 553–558. [CrossRef]
7. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Martín-Antón, L.J.; González-Diez, I.; Carbonero-Martín, I. Using eye tracking technology to analyse cognitive load in multichannel activities in university students. *Int. J. Hum.-Comput. Interact.* **2023**, *40*, 3263–3328. [CrossRef]
8. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* **1998**, *12*, 257–285. [CrossRef]

9. Sweller, J.; Chandler, P.; Tierney, P.; Cooper, M. Cognitive load as a factor in the structuring of technical material. *J. Exp. Psychol. Gen.* **1990**, *119*, 176–192. [CrossRef]
10. Mayer, R.E. *Multimedia Learning*; Cambridge University Press: New York, NY, USA, 2001. [CrossRef]
11. Tindall-Ford, S.; Chandler, P.; Sweller, J. When two sensory modes are better than one. *J. Exp. Psychol. Appl.* **1997**, *3*, 257–287. [CrossRef]
12. Baddeley, A. Working memory. *Science* **1992**, *255*, 556–559. [CrossRef]
13. Schnotz, W.; Kürschner, C.A. Reconsideration of Cognitive Load Theory. *Educ. Psychol. Rev.* **2007**, *19*, 469–508. [CrossRef]
14. Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef]
15. Peterson, L.; Peterson, M. Short-term retention of individual verbal items. *J. Exp. Psychol.* **1959**, *58*, 193–198. [CrossRef]
16. Marcus, N.; Cooper, M.; Sweller, J. Understanding instructions. *J. Educ. Psychol.* **1996**, *88*, 49–63. [CrossRef]
17. Vygotski, L.S. Learning and mental development at school age. In *Educational Psychology in the U.S.S.R.*; Simon, B.J., Simon, J., Eds.; Routledge & Kegan Paul: London, UK, 1963; pp. 21–34.
18. Sweller, J. Cognitive load theory and individual differences. *Learn. Individ. Differ.* **2024**, *110*, 102423. [CrossRef]
19. Ahmad, M.I.; Keller, I.; Robb, D.A.; Lohan, K.S. A framework to estimate cognitive load using physiological data. *Pers. Ubiquitous Comput.* **2023**, *27*, 2027–2041. [CrossRef]
20. Lee, Y.; Szulewski, A.; Young, J.Q.; Donkers, J.; Jarodzka, H.; Van Merriënboer, J.J.G. The medical pause: Importance, processes, and training. *Med. Educ.* **2021**, *55*, 1152–1160. [CrossRef]
21. Seufert, T. The interplay between self-regulation in learning and cognitive load. *Educ. Res. Rev.* **2018**, *24*, 116–129. [CrossRef]
22. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; Weijer, J.V.D. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; Oxford University Press: Oxford, UK, 2011.
23. Scheutz, M.; Aeron, S.; Aygun, A.; de Ruiter, J.P.; Fantini, S.; Fernandez, C.; Haga, Z.; Nguyen, T.; Lyu, B. Estimating Systemic Cognitive States from a Mixture of Physiological and Brain Signals. *Top. Cogn. Sci.* **2024**, *16*, 485–526. [CrossRef]
24. Mayer, R.E. Unique contributions of eye-tracking research to the study of learning with graphics. *Learn. Instr.* **2010**, *20*, 167–171. [CrossRef]
25. Mutlu-Bayraktar, D.; Ozel, P.; Altindis, F.; Yilmaz, B. Relationship between objective and subjective cognitive load measurements in multimedia learning. *Interact. Learn. Environ.* **2020**, *31*, 1322–1334. [CrossRef]
26. Koć-Januchta, M.; Höffler, T.; Thoma, G.-B.; Prechtl, H.; Leutner, D. Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures—An eye-tracking study. *Comput. Hum. Behav.* **2017**, *68*, 170–179. [CrossRef]
27. Duchowski, A.T.; Krejtz, K.; Gehrer, N.A.; Bafna, T.; Bækgaard, P. The low/high index of pupillary activity. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
28. Critchley, H.D. Electrodermal responses: What happens in the brain. *Neuroscientist* **2002**, *8*, 132–142. [CrossRef] [PubMed]
29. Winter, M.; Pryss, R.; Probst, T.; Reichert, M. Towards the applicability of measuring the electrodermal activity in the context of process model comprehension: Feasibility study. *Sensors* **2020**, *20*, 4561. [CrossRef] [PubMed]
30. Kramer, A.F. Physiological metrics of mental workload. In *Multiple Task Performance*; CRC Press: Boca Raton, FL, USA, 1991; pp. 279–328.
31. Abbad-Andaloussi, A.; Burattin, A.; Slaats, T.; Kindler, E.; Weber, B. Complexity in declarative process models: Metrics and multi-modal assessment of cognitive load. *Expert Syst. Appl.* **2023**, *233*, 120924. [CrossRef]
32. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Escolar-Llamazares, M.C.; González-Díez, I.; Martín Antón, L.J. Using integrated multimodal technology: A way to personalised learning in Health Sciences and Biomedical engineering Students. *Appl. Sci.* **2024**, *14*, 7017. [CrossRef]
33. Liberman, L.; Dubovi, I. The effect of the modality principle to support learning with virtual reality: An eye-tracking and electrodermal activity study. *J. Comput. Assist. Learn.* **2023**, *39*, 547–557. [CrossRef]
34. Li, Q.; Luximon, Y.; Zhang, J.; Song, Y. Measuring and classifying students' cognitive load in pen-based mobile learning using handwriting, touch gestural and eye-tracking data. *Br. J. Educ. Technol.* **2024**, *55*, 625–653. [CrossRef]
35. Raftery, A.E. Use and communication of probabilistic forecasts. Statistical Analysis and Data Mining: The ASA. *Data Sci. J.* **2016**, *9*, 397–410. [CrossRef]
36. Sáiz-Manzanares, M.C. Metacognition and Artificial Intelligence: Beyond the Parallelism of Functioning [Metacognición e Inteligencia Artificial: Más allá del Paralelismo de Funcionamiento]. Ph.D. Thesis, University of Burgos, Burgos, Spain, 2019. Available online: http://hdl.handle.net/10259/5357 (accessed on 12 August 2024).
37. Gogna, Y.; Tiwari, S.; Singla, R. Evaluating the performance of the cognitive workload model with subjective endorsement in addition to EEG. *Med. Biol. Eng. Comput.* **2024**, *62*, 2019–2036. [CrossRef]
38. Gogna, J.; Tiwari, S.; Singla, R. Mental workload assessment of gamers' eeg with multi-domain feature-based cognitive model and its validation. *Biomed. Eng. Appl. Basis Commun.* **2024**, *36*, 2450022. [CrossRef]
39. Sazuka, N.; Katsumata, K.; Komoriya, Y.; Oba, T.; Ohira, H. Association of brain-autonomic activities and task accuracy under cognitive load: A pilot study using electroencephalogram, autonomic activity measurements, and arousal level estimated by machine learning. *Front. Hum. Neurosci.* **2024**, *18*, 1272121. [CrossRef]
40. Minissi, M.E.; Altozano, A.; Marín-Morales, J.; Giglioli, I.A.C.; Mantovani, F.; Alcañiz, M. Biosignal comparison for autism assessment using machine learning models and virtual reality. *Comput. Biol. Med.* **2024**, *171*, 108194. [CrossRef] [PubMed]

41. Karmakar, S.; Kamilya, S.; Dey, P.; Guhathakurta, P.K.; Dalui, T.M.; Bera, T.K.; Halder, S.; Koley, C.; Pal, T.; Basu, A. Real time detection of cognitive load using fNIRS: A deep learning approach. *Biomed. Signal Process. Control* **2023**, *80*, 104227. [CrossRef]
42. Shaposhnyk, O.; Yanushkevich, S.; Babenko, V.; Chernykh, M.; Nastenko, I. Inferring Cognitive Load Level from Physiological and Personality Traits. In Proceedings of the 2023 International Conference on Information and Digital Technologies (IDT), Zilina, Slovakia, 20–22 June 2023; pp. 233–242. [CrossRef]
43. Murdaca, G.; Banchero, S.; Casciaro, M.; Tonacci, A.; Billeci, L.; Nencioni, A.; Pioggia, G.; Genovese, S.; Monacelli, F.; Gangemi, S. Potential Predictors for Cognitive Decline in Vascular Dementia: A Machine Learning Analysis. *Processes* **2022**, *10*, 2088. [CrossRef]
44. Sáiz-Manzanares, M.C.; Escolar-Llamazares, M.-C.; Arnaiz González, Á. Effectiveness of Blended Learning in Nursing Education. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1589. [CrossRef]
45. Sáiz-Manzanares, M.C.; Ramos Pérez, I.; Arnaiz-Rodríguez, Á.; Rodríguez-Arribas, S.; Almeida, L.; Martin, C.F. Analysis of the learning process through eye tracking technology and feature selection techniques. *Appl. Sci.* **2021**, *11*, 6157. [CrossRef]
46. Sáiz-Manzanares, M.C.; Rodríguez-Díez, J.J.; Marticorena, R.; Zaparaín, M.J.; Cerezo, R. Lifelong Learning from Sustainable Education: An Analysis with Eye Tracking and Data Mining Techniques. *Sustainability* **2020**, *12*, 1970. [CrossRef]
47. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Martín-Antón, L.J.; Almeida, L.; Carbonero-Martín, I. Application and challenges of eye tracking technology in Higher Education. *Comunicar* **2023**, *76*, 35–46. [CrossRef]
48. Langner, M.; Toreini, P.; Maedche, A. Cognitive state detection with eye tracking in the field: An experience sampling study and its lessons learned. *i-com* **2024**, *23*, 109–129. [CrossRef]
49. Wang, A.; Huang, C.; Wang, J.; He, D. The association between physiological and eye-tracking metrics and cognitive load in drivers: A meta-analysis. *Transp. Res. Part F Psychol. Behav.* **2024**, *104*, 474–487. [CrossRef]
50. Foltyn, A.; Deuschel, J.; Lang-Richter, N.R.; Holzer, N.; Oppelt, M.P. Evaluating the robustness of multimodal task load estimation models. *Front. Comput. Sci.* **2024**, *6*, 1371181. [CrossRef]
51. Havugimana, F.; Moinudin, K.A.; Yeasin, M. Deep Learning Framework for Modeling Cognitive Load From Small and Noisy EEG Data. *IEEE Trans. Cogn. Dev. Syst.* **2024**, *16*, 1006–1015. [CrossRef]
52. Khan, M.K.; Asadi, H.; Zhang, L.; Qazani, M.R.C.; Oladazimi, S.; Loo, C.K.; Lim, C.P.; Nahavandi, S. Application of artificial intelligence in cognitive load analysis using functional near-infrared spectroscopy: A systematic review. *Expert Syst. Appl.* **2024**, *249*, 123717. [CrossRef]
53. Hijazi, H.; Gomes, M.; Castelhano, J.; Castelo-Branco, M.; Praça, I.; de Carvalho, P.; Madeira, H. Dynamically predicting comprehension difficulties through physiological data and intelligent wearables. *Sci. Rep.* **2024**, *14*, 13678. [CrossRef] [PubMed]
54. Gong, J.; Cao, S.; Korivand, S.; Jalili, N. Reconstructing human gaze behavior from EEG using inverse reinforcement learning. *Smart Health* **2024**, *32*, 100480. [CrossRef]
55. Yang, H.; Wu, J.; Hu, A.; Lv, C. Real-Time Driver Cognitive Workload Recognition: Attention-Enabled Learning With Multimodal Information Fusion. *IEEE Trans. Ind. Electron.* **2024**, *71*, 4999–5009. [CrossRef]
56. Chango, W.; Lara, J.A.; Cerezo, R.; Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. Wiley Interdiscip. *Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1458. [CrossRef]
57. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
58. Ozdel, S.; Rong, Y.; Albaba, B.M.; Kuo, Y.-L.; Wang, X.; Kasneci, E. A Transformer-Based Model for the Prediction of Human Gaze Behavior on Videos. *arXiv* **2024**, arXiv:2404.07351v1. [CrossRef]
59. Toy, J.; MacAdam, J.; Tabor, P. Metacognition is all you need? Using Introspection in Generative Agents to Improve Goal-directed Behavior. *arXiv* **2024**, arXiv:2401.10910. [CrossRef]
60. An, Y.; Zhang, K.; Chai, Y.; Zhu, Z.; Liu, Q. Gaussian Mixture Variational-Based Transformer Domain Adaptation Fault Diagnosis Method and Its Application in Bearing Fault Diagnosis. *IEEE Trans. Ind. Inform.* **2024**, *20*, 615–625. [CrossRef]
61. An, Y.; Li, Z.; Li, Y.; Zhang, K.; Zhu, Z.; Chai, Y. Few-Shot Learning-Based Fault Diagnosis Using Prototypical Contrastive-Based Domain Adaptation Under Variable Working Conditions. *IEEE Sens. J.* **2024**, *24*, 25019–25029. [CrossRef]
62. Suzuki, Y.; Wild, F.; Scanlon, E. Measuring cognitive load in augmented reality with physiological methods: A systematic review. *J. Comput. Assist. Learn.* **2024**, *40*, 375–393. [CrossRef]
63. Román Sánchez, J.M.; Gallego Rico, S. *ACRA Escalas de Estrategias de Aprendizaje [Learning Strategy Scales]*; TEA: Madrid, Spain, 2008; Available online: http://www.web.teaediciones.com/Ejemplos/ACRA_extracto_web.pdf (accessed on 21 October 2024).
64. Eibe, F.; Hall, M.A.; Witten, I.H. The WEKA Workbench. In *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA; San Francisco, CA, USA, 2016.
65. IBM Corp. *SPSS Statistical Package for the Social Sciences (SPSS)*, 28th ed.; IBM Corp.: Armonk, NY, USA, 2024.
66. IBM Corp. *Structural Equation Modeling (AMOS SPSS)*, 29th ed.; IBM Corp.: Armonk, NY, USA, 2024.
67. Cohen, J.; Cohen, P.; West, G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed.; Routledge: New York, NY, USA, 2002. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.