aerospace

MDPI

Article

Mental Workload as a Predictor of ATCO's Performance: Lessons Learnt from ATM Task-Related Experiments

Enrique Muñoz-de-Escalona [1,*], Maria Chiara Leva [1] and José Juan Cañas [2]

[1] Environmental Sustainability and Health Institute, Technological University Dublin, D07 EWV4 Dublin, Ireland; mariachiara.leva@tudublin.ie
[2] Mind, Brain and Behaviour Research Centre, University of Granada, 18971 Granada, Spain; delagado@ugr.es
* Correspondence: enrique.munozdeescalonafernandez@tudublin.ie

Abstract: Air Traffic Controllers' (ATCos) mental workload is likely to remain the specific greatest functional limitation on the capacity of the Air Traffic Management (ATM) system. Developing computational models to monitor mental workload and task complexity is essential for enabling ATCOs and ATM systems to adapt to varying task demands. Most methodologies have computed task complexity based on basic parameters such as air-traffic density; however, literature research has shown that it also depends on many other factors. In this paper, we present a study in which we explored the possibility of predicting task complexity and performance through mental workload measurements of participants performing an ATM task in an air-traffic control simulator. Our findings suggest that mental workload measurements better predict poor performance and high task complexity peaks than other established factors. This underscores their potential for research into how different ATM factors affect task complexity. Understanding the role and the weight of these factors in the overall task complexity confronted by ATCos constitutes one of the biggest challenges currently faced by the ATM sphere and would significantly contribute to the safety of our sky.

Keywords: mental workload; task complexity; performance prediction; workload measures; latency differences; dissociations

Citation: Muñoz-de-Escalona, E.; Leva, M.C.; Cañas, J.J. Mental Workload as a Predictor of ATCO's Performance: Lessons Learnt from ATM Task-Related Experiments. *Aerospace* **2024**, *11*, 691. https://doi.org/10.3390/aerospace11080691

Academic Editor: Eri Itoh

Received: 9 June 2024
Revised: 16 August 2024
Accepted: 20 August 2024
Published: 22 August 2024

1. Introduction

The Air Traffic Control (ATC) task is a dynamic and complex task in which performance depends on Air Traffic Controllers' (ATCo) skills acquired through learning, as well as on psychological factors, such as emotions, stress, mental workload and fatigue. Among all these psychological factors, mental workload is considered the most important factor in predicting how complex the ATC task is and how the ATCo will perform throughout task development [1–3]. For example, Edwards et al. (2017) found that mental workload, along with situational awareness, can have a compounded negative impact on the ATCos' performance [4]. Another study from the same author suggests that this relationship may be influenced by context, level of automation and the direction in which mental workload changes over time (whether it transitions from low to high or high to low) [5]. In a recent study, Alaydi and Ng (2024) found a negative relationship between workload and ATCos' performance in the ATM context [6]. Similarly, Balta et al. (2024) evidenced that the performance of ATCos fluctuates with changes in their perception of time, influenced by varying levels of mental workload; when experiencing a high mental workload, ATCos perceive time as passing more quickly than it actually does, leading to more reactive behavior, which increases the likelihood of errors and decreases overall performance [7].

Nevertheless, there has always been a significant interest in the concept of "task complexity" in the ATM field, justified by the notion that the effectiveness, efficiency, and safety of tasks performed by ATCos are fundamentally influenced by task complexity, which drives mental workload [3]. The ATM tasks performed by ATCos are multifactorial,

dynamic, symbolic and uncertain [8]. ATC task-load factors are variable and challenging to predict. Traffic factors (e.g., traffic density) and operational constraints (e.g., restrictions on available airspace) are inherently dynamic. Additionally, airspace factors (e.g., number of available flight levels), infrastructure issues (e.g., closed taxiways), technological resources (e.g., communication breakdowns), and workforce management (e.g., illnesses or substitutions) are subject to change, requiring ATCos to continuously adapt to a complex and evolving environment. Current studies aim to find methodologies to assess task complexity in ATM so that ATCos and ATM systems can adapt as effectively as possible to the anticipated task complexity. It is intended that this process of adaptation is carried out through airspace restructuration to reduce predicted complexity and by training the ATCos in the necessary skills to successfully face changes in task complexity. For this reason, researchers are working on computational models for predicting task complexity, such as COMETA (COgnitive ModEl for aTco workload Assessment), which is a computational model designed to monitor and predict the mental workload of the Air Traffic Controller through a computerized model that can be exploited in simulation or real environments for monitoring and predicting task complexity [9].

Task complexity in the ATM field has been directly computed by several environmental parameters [10]. Traditionally, most formulas have computed ATM task complexity based on simple parameters such as air-traffic density; however, literature research has shown that task complexity depends not only on the gross amount of aircrafts shown in the radar screen but on so many other factors that affect the configuration and evolution of different flights, modelling and ultimately, the cognitive complexity of the scenario, i.e., the mental workload supported by ATCos [9]. Considering that mental workload can be assessed through different methodologies and it predicts task complexity and performance, it could be possible to use different mental workload parameters in order to derive salient characteristics of the ATM environment that determines task complexity.

In this paper, we aim to explore whether by monitoring subjective and psychophysiological measurements traditionally used for assessing mental workload levels, if it would be possible to predict task complexity and performance, so that we could ultimately identify, with further research, what aspects of a scenario or a task in an ATM context may foster higher task complexity and therefore performance impairment. However, there is one issue that must be considered, and it is the fact that mental workload assessment has several limitations, mainly because of the lack of convergence that we can observe between the different methodologies for assessing it. Therefore, it is also useful to study latency differences between various mental workload measurements that would ultimately affect the prediction of task complexity and performance from mental workload indications.

2. Workload and Human Performance Assessment: A Quick Overview of the State of the Art

Mental workload is an abstract concept that is highly relevant in the ergonomics and work sciences spheres and we could state, in general terms, that mental workload is gaining more and more importance in our modern world. Managing mental workload is essential both to guarantee occupational safety and health [11,12] and to optimize and increase business productivity [13]. Even in today's world in which AI and automation are being rapidly developed to reduce cognitive load and effort, having robust and valid methodologies for assessing mental workload and related factors such as mental fatigue remains fundamental. Performance problems can arise not only in high mental workload scenarios but also in low workload situations, which can lead to risky conditions such as the "out of the loop" phenomenon [14,15]. Moreover, while technological advances aim to reduce mental workload and fatigue, these tools still require validation, which necessitates the ability to measure these factors in a valid a reliable way. These are the main reasons for the development of multiple cognitive models in recent years, which would allow predicting human mental workload relying on the interaction of

particular additive factors [16–18]. After all, computational models have proven to be adequate, very effective and reliable for computing predictions of scientific models in many domains [19]. However, for developing and validating computational models of mental workload, as well as for simply making research possible on this construct, we first need to be able to assess mental workload in a proper and reliable way. This is the main reason for the exponential increase in the number of studies addressing mental workload measurement methodologies and techniques [20–25]. Nevertheless, the key problem both for developing mental workload research and for managing mental workload in applied environments is that its assessment has proven to be problematic and inaccurate.

Mental workload is a psychological construct and that means that it is intangible and cannot be measured directly; however, literature research has shown that there are different indexes that are capable, in an indirect manner, to reflect variations in the magnitude of this construct [26]. In particular, it is well established that mental workload changes can be reflected by different indirect indexes that are known as the "three primary reflections" of mental workload, which are as follows: (1) performance measures, (2) psychophysiological responses and (3) subjective reports of mental workload [27,28].

Hence, if we consider that these three indexes are supposed to reflect the changes that occur in the magnitude of the mental workload construct, we would expect to find a high degree of convergence (high correlations) between them. This means that we should be able to assess this construct with each primary measure in an independent manner and if we compare the evolution of those measurements over time, we should find a high degree of coincidence between those measurements in reflecting mental workload variations (this is what has been defined as convergence between measures). For example, imagine that a task carried out by an employee suddenly becomes harder; we would expect to find a decrease in that employee's performance, in conjunction with an increase in his/her psychophysiological responses to face the higher complexity of the task (note that other psychophysiological responses such as heart rate variability [29] and blink rate [30], have been shown to decrease during periods of high workload) and an increase in his/her perceived amount of mental workload. However, what we have seen both in literature research and in applied contexts is that this logical sequence of mental workload reflections does not always show such a clear pattern. In other words, the emergence of dissociation (inconsistencies) between measures is a phenomenon much more frequent than we might think [31–33]. This fact is a major issue for our basic research and mainly for our practice, and the human factors and ergonomics science community must untangle the human mental workload assessment problem. In 2017, Hancock [31] reported that divergences between measures might be caused by several different factors, among which temporal differences between measures appeared to be a possible affecting factor. We also think that temporal differences between the different methodologies for assessing mental workload may partly explain, as Hancock reports, certain inconsistencies between mental workload measures.

3. Experimental Design: Methods and Materials

In this study, participants had to perform an ATM simulation experiment for 120 min. Participants were trained to use the ATM simulator and they were instructed to avoid possible conflicts (between flights) on the radar screen. Task complexity and time on task (TOT) were our manipulated variables and we registered the following three different mental workload primary measures: conflict rate as our performance measure, pupil size as our psychophysiological measure and Instantaneous Self-Assessment (ISA) scale as our subjective measure. The study was divided into two experimental conditions. During experimental condition 1, we established 5 min intervals, so pupil size and performance measures were averaged for each interval, whereas the subjective mental workload measure was obtained as a discrete value at the end of each interval. In experimental condition

2, we aimed to increase the granularity of collected data by reducing time intervals to 2 min.

We aimed to test the following hypotheses:

Hypothesis 1 (H1). *It is possible to use the ISA scale as a predictor of operator performance and task complexity.*

Hypothesis 2 (H2). *It is possible to use pupillometry as a predictor of operator performance and task complexity.*

Hypothesis 3 (H3). *Different latencies exist among mental workload measurements in reflecting mental workload variations over time.*

3.1. Materials and Instruments

3.1.1. ^{ATC}Lab-Advanced Software

We used a free licence software called ^{ATC}Lab-Advanced for simulating different Air Traffic Control (ATCo) scenarios that were carried out by participants during the training and the data collection stage [34]. This software is very realistic and very easy to use by participants, who can learn to use the simulator in just a few training sessions. It has been employed in various research areas, such as intelligence and working memory (e.g., [35]), motivation and decision-making (e.g., [36]), emotion (e.g., [37]), performance (e.g., [38]), and mental workload assessment in the ATM domain (e.g., [39]). Its extensive use across different fields demonstrates its validity and reliability for simulating real-world ATM scenarios. Additionally, it has been utilized in ATM research to develop a computer-based aerodrome control tower simulator for ATC [40]. On the other hand, ^{ATC}Lab-Advanced provides high experimental control, as most air-traffic parameters can be modified according to the objectives of the research. More specifically, we can modify different structural and dynamic parameters of the scenario, such as aircraft pathways and the size of the scenario; and the air-traffic density and flight parameters (speed, altitude, route, etc.), respectively. Participants were taught, during the learning stage, to use the different air-traffic management tools provided by the simulator (distance scale, speed and altitude change tools) to keep flights far enough from each other in order to guarantee the aviation safety. Finally, ^{ATC}Lab-Advanced software provides a .log recorded file with the performance data of participants after performing the executed scenario.

3.1.2. Scenarios and Indicators of Task Complexity

In this study, we used different scenarios depending on whether participants were performing the training or the data collection stage. During the learning stage, we used default scenarios provided by the developers of the simulator, but during the latter stage, we programmed a specific scenario for our experimental purpose. The characteristics of this particular scenario were the following:

1. The scenario was programmed with the general purpose of producing task demand variations with time on task (TOT). In other words, the amount of mental workload experienced by participants was different throughout the execution of the task in order to collect the variations caused in the different mental workload indexes and see how they would affect operator performance.
2. Nine initial aircrafts were presented to participants, six of which were controlled initially by them.
3. Along the execution of the scenario, participants controlled a total amount of 70 aircrafts; 50 of them travelled from external (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) to internal locations (11, 12) and, conversely, 20 travelled from internal to external locations (Figure 1).

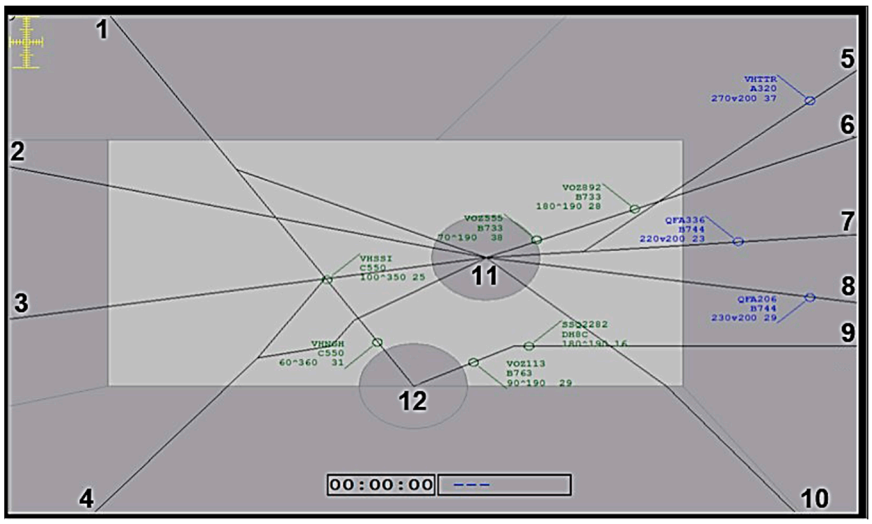


Figure 1. ATCLab-Advanced initial scenario screen during data collection stage. Outbound air traffic is displayed in green, while inbound air traffic in blue.

3.1.3. Tobii T120 Eye-Tracker

The Tobii T120 (Tobii Technology AB, Stockholm, Sweden) is a static eye-tracker system originally designed by the "Tobii Video System" to collect and record gaze behaviour and ocular parameters of participants. The system is able to record data with a sampling rate of 120 Hz and stand out for its high precision, low intrusiveness and excellent quality in data collection. After a quick calibration procedure, participants can use the eye-tracker in a natural way, with freedom of head movements and relaxed position.

3.1.4. Instantaneous Self-Assessment Scale

The Instantaneous Self-Assessment Scale is a subjective mental workload scale that allows the assessment of subjective perceived mental workload of participants while performing their task (Figure 2). Its primary advantage is that it enables continuous, real-time assessment of subjective mental workload (an online method), unlike most mental workload questionnaires and scales that can only be administered after task completion (an offline method) and therefore depend on participants' memory. This easy handling scale was designed for assessing subjective mental workload in the ATM field, but it can be applied in most fields. The scale has a range from 1 (absence of mental workload) to 5 (highest perceived mental workload). Participants must assess their perceived mental workload at specific time intervals established by the researcher, which vary depending on the experimental condition considered. While this method has the issue of disrupting participants' main tasks, research in the literature indicates that it is considered one of the least intrusive online methods for evaluating mental workload [41,42].

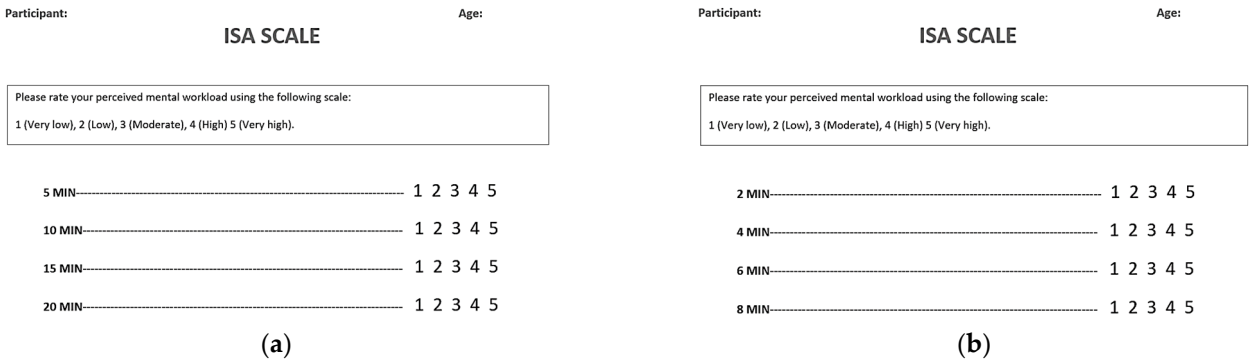


Figure 2. (**a**) ISA scale, 5 min intervals; (**b**) ISA scale, 2 min intervals.

3.2. Experimental Conditions and Participants

The authors designed two identical experimental conditions, including one where the ISA values were collected every 5 min and one in which the values were collected every 2 min. We received advice from the Spanish Air Traffic Management Research Center (CRIDA A.I.E), indicating that acquiring ISA values in real environments at intervals shorter than 5 min could be highly intrusive for ATCOs and negatively impact their MW and performance. However, we also decided to gather ISA values at a 2 min interval to determine if higher granularity could provide better insights into the correlation among the following three primary mental workload measures: conflict rate as the performance measure, pupil size as the psychophysiological measure and ISA scale as the subjective measure.

The participants of the first experimental condition were 32 undergraduate psychology students at Granada University (71.88% female; age range = 19–29; average age = 22.1 and median = 22). The participants of the second experimental condition were 26 undergraduate psychology students at the same university (76.92% female; age range = 18–39; average age = 22.1 and median = 21) and they were also rewarded with extra credit. The participant requirements were as follows: (1) no prior experience with the ^{ATC}Lab-Advanced software for ATM tasks, (2) Spanish as a native language, and (3) visual acuity or corrected vision using contact lenses, as glasses interfere with the eye-tracking device used for data collection. The study involved psychology students as they were easily recruited by offering two experimental vouchers that granted extra credit as an incentive for their participation. The lack of experience among selected participants in ATM tasks ensured that all participants started from the same baseline of expertise as novices, preventing different prior levels of experience from biasing the results. Additionally, the participants' lack of experience likely increased the overall level of mental workload experienced during task development, especially during high task-demand peaks. This was beneficial for our study, as we needed to induce high mental workload peaks. Recruitment was conducted through posters and flyers distributed around the university and an advertisement on the university's online platform for experiments. It is worth noting that most participants were female, reflecting the predominantly female student population in the psychology degree at the University of Granada. This study was developed according to the recommendations of the local ethical guidelines of the committee of Granada University named "Comité de Ética de Investigación Humana". Participants in both experiments gave written informed consent in accordance with the Declaration of Helsinki.

3.3. Procedure

Participants were instructed to use the ^{ATC}Lab-Advanced Software to perform the experimental ATCo task developed for the study. Thus, the procedure was divided into two different stages, including one learning stage and a subsequent data collection stage during the following day, which are outlined below:

1. Learning stage: This first stage lasted 90 min and the main goal was to achieve proper learning and training of the ATM simulator by participants so that they could handle it comfortably before the data collection stage. First, participants had to read and sign the informed consent document and read a brief manual about the use of the simulator. Then, the participants were informed about their task goal (maintain aviation safety by preventing eventual conflicts between aircrafts) and the manual was reviewed in detail between participants and the researcher to guarantee proper understanding of the simulator use. Finally, participants trained themselves with the simulator by performing 6 training scenarios that were executed in order of difficulty. Participants could always ask questions to the researcher and consult the manual if necessary. At the end of the learning stage, the researcher checked the performance of participants to ensure their learning.
2. Data collection stage: This stage took place during the day after the training stage. It lasted 120 min and the main goal was to collect empirical data of the participants during the execution of the experimental ATM scenario. More specifically, we collected

performance (no. of conflicts between aircrafts) and psychophysiological (pupil size) objective data, as well as subjective mental workload reports (ISA scale). The procedure of the data collection stage for each participant was as follows: first, the researcher calibrated the eye-tracker system and told the participant to avoid head movements during the execution of the task. Then, the participant was instructed to report perceived mental workload periodically, every 5 min (during experimental condition 1), or every 2 min (during experimental condition 2). A periodic alarm would sound to advise the right time to record the ISA scale. Finally, the participant would start performing the experimental scenario for 120 min while recording. Once the experimental session finished, the participants were thanked for their participation and awarded with credits.

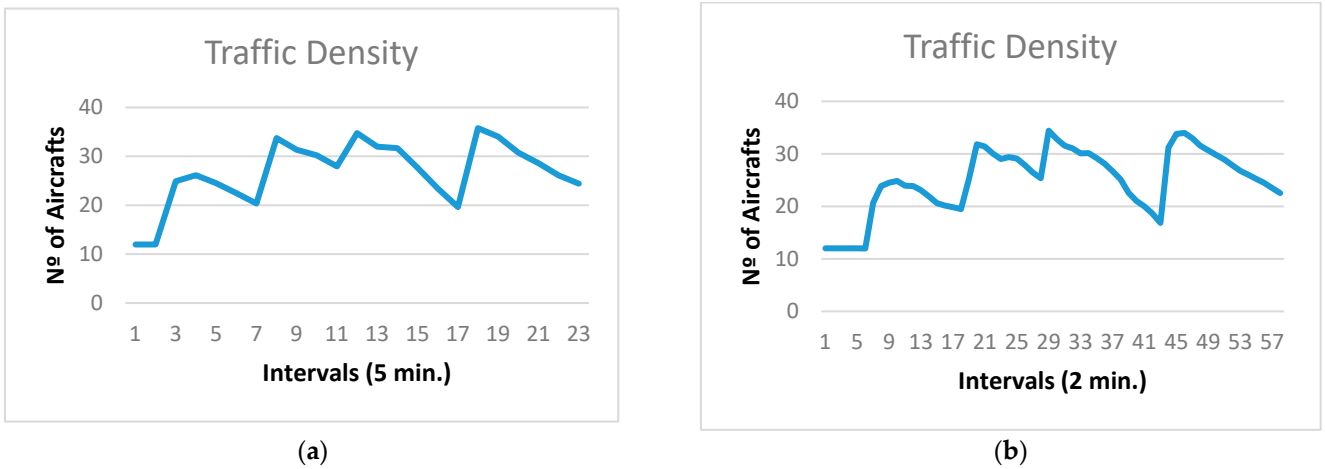
3.4. Experimental Room Conditions

During the training stage, no particular room conditions were followed and participants could work in one of three different training rooms equipped with everything needed for learning how to use the ^{ATC}Lab-Advanced simulator. Conversely, the data collection stage required a standardization of room conditions in order to minimize the possible impact of ambient conditions in data recordings. Hence, temperature was kept constant to 21 °C and screen distance to participants was kept at 60 cm. Furthermore, we took special care to maintain consistent lighting conditions, preventing natural light from entering the room and ensuring uniform artificial lighting and screen brightness. This was crucial because pupil size can react to even minor variations in illumination, including differences across different areas of the screen based on where the participant looks. However, the ^{ATC}Lab-Advanced scenario screen was kept unchanged and displayed in full screen, with any illumination differences across the screen being negligible.

3.5. Variables

3.5.1. Independent Variables

Task Complexity: One of the key factors affecting mental workload is task complexity, as a higher task complexity will increase operators' mental workload. In this study, we adjusted the air-traffic density (no. of aircrafts on the radar screen) to manipulate task complexity and thus mental workload. In Figure 3, we can appreciate the air-traffic density manipulation carried out throughout intervals (120 min overall) in both experimental conditions.



(**a**) (**b**)

Figure 3. (**a**) Air-traffic density through intervals for experimental condition 1; (**b**) air-traffic density through intervals for experimental condition 2.

Time on task: The total duration of the experimental task was 120 min. In experimental condition 1, we established 24 intervals of 5 min each, whereas in experimental condition 2, we implemented 60 intervals of 2 min each.

3.5.2. Dependent Variables

Performance: ATCos' performance can be assessed by several different indexes; however, we decided to use the conflict rate as it largely correlates with task complexity [43]. The conflict rate is the result of dividing the no. of conflicts by the no. or aircrafts present on the radar screen. We must note that considering just the no. of conflicts as our performance index would bias our data, as air-traffic density is directly related to the number of conflicts between aircrafts.

Pupil size: Mental workload can be reflected by several different psychophysiological responses such as the extrapolation of features from EEG signals [44], Heart Rate Variability (HRV) [45], psychogalvanic response [46] and several ocular metrics [47], among others. It is worth noting that in certain fields, such as ATM, there has been significant interest in the use of ocular metrics due to the specific need for non-invasive systems to monitor mental workload and fatigue in real-time. Eye behaviour metrics can be estimated to reflect fatigue, drowsiness, visual distraction, cognitive distraction, and workload (e.g., [47,48]). Static eye-tracking systems allow monitoring and recording in real time the entire set of ocular parameters that comprise visual behaviour in an unobtrusive way. Literature research has shown that there are several different ocular metrics that can be used for estimating mental workload, such as gaze patterns [49], blink rate [50] and pupil size [51–60]. Research has traditionally linked mental effort with pupil size variations. In the 1960s, Hess explored pupillary changes associated with cognitive effort. Specifically, he observed that pupil size increased as the experimenter increased the difficulty of multiplications [61]. During the 1960s and 1970s, other researchers also reached the same conclusions through different experimental works, including recall tasks [62,63], mental calculation [64–66] and continuous processing tasks [67]. However, it was Kahneman's (1973) studies on pupillary changes and cognitive activity that had the greatest impact in the field of psychology because they were integrated into his well-known "attentional theory". He found a strong relationship between pupil dilation and cognitive processing, which is why Kahneman (1973) defends this psychophysiological measure as a relevant indicator of the attentional resources involved in the task [68]. In conclusion, and in light of the data collected in decades of research, it seems that pupil diameter increase is a good indicator of mental workload.

According to Beatty (1982), pupil diameter is a continuous variable expressed over time, with responses taking the following two forms: a tonic response, where the pupil changes size for a sustained period, and a phasic response, which is an event-related transient change [69]. In cognitive science research, the focus is typically on evaluating phasic fluctuations of this variable in response to time-locked events (e.g., the onset of a picture or sound, or specific discrete events at precise times). This is known as a task-evoked pupillary response (TEPR) [69]. TEPR studies have shown that pupil size can react to mental workload within approximately 200 milliseconds to 9 s (e.g., [70]). However, our study cannot be directly compared to TEPR studies since we do not present discrete stimuli. Instead, we use a simulated ATM scenario involving complex interactions among various stimuli, leading to an overall shift in task demands driven by increased traffic density. Therefore, in this study, we analysed the tonic response of pupil size to task demands by averaging values over regular intervals. In experimental condition 1, 24 intervals of 5 min durations were set, while in experimental condition 2, there were 60 intervals of 2 min durations. However, following Sebastiaan Mathôd's recommendations [71], a baseline correction was performed to prevent the slow and irregular variations that characterise the absolute values of pupil diameter. This procedure was carried out as follows: in experimental condition 1, we used the first interval average pupil size value as a baseline (5 min) to which we subtracted the remaining 23 intervals. In an equivalent manner, in experimental condition 2, we took the first two intervals' average pupil size value as a

baseline (4 min) to which we subtracted the remaining 57 intervals. We performed analyses both for left and right eye pupils. We must note that individual differences in pupil size were addressed statistically (repeated measures ANOVA). The main drawback of this methodology is that pupil size is primarily influenced by lighting conditions, which is why it is crucial to maintain constant lighting conditions when using pupil size to assess mental workload. While this is manageable in laboratory settings where lighting and brightness are controlled, it becomes challenging under real-world conditions where consistent lighting cannot be guaranteed.

Subjective mental workload: Traditional subjective mental workload scales and questionnaires can only be implemented after the execution of the task; however, we needed to obtain a real-time assessment of the mental workload of participants in order to allow us to make correlations between the three mental workload indexes. For that reason, we decided to use the ISA scale, which can easily be handled and modified to give momentary perceived mental workload ratings at certain time intervals. In experimental condition 1, we obtained ISA ratings at 5 min intervals, while in experimental condition 2, we obtained ratings at 2 min intervals. Hence, we obtained 24 intervals in experimental condition 1 and 60 intervals in experimental condition 2, from which we discarded the first interval in experimental condition 1, and the first 2 intervals in experimental condition 2; this gave us 23 remaining intervals for data analysis in experimental condition 1, and 58 intervals in experimental condition 2 (see Figures 2 and 3).

4. Overview of Key Results

4.1. Experimental Condition 1

The results were analyzed using a one-way within-subjects ANOVA. Please refer to Figure 4 for the obtained results.

First, considering the performance results, a significant effect of the intervals was found, where $F(22,682) = 42.44$, $MSe = 0.001$ and $p < 0.001$. This effect was mainly due to two poor performance peaks at intervals 8 and 18. However, during the other intervals, the participants' performance was adequate. These two poor performance intervals coincided with two peaks in high air-traffic density, indicating higher task complexity.

Regarding pupil size, a significant effect of intervals was found for both the right and left pupils, where $F(22,682) = 8.34$, $MSe = 0.005$ and $p < 0.001$, and $F(22,682) = 7.98$, $MSe = 0.005$ and $p < 0.001$, respectively. Pupil size reached maximum values during both poor performance peaks, particularly from intervals 9 to 12 and 19 to 21, reflecting changes in mental workload over time.

As for subjective reports of mental workload, a significant effect of intervals was again found, where $F(22,682) = 25.17$, $MSe = 0.758$ and $p < 0.001$. The graph shows that subjective mental workload fluctuated over time, reaching maximum values during the poor performance peaks, similar to pupil size, specifically during intervals 8 to 12 and 18 to 21.

Regarding the correlations between the different measures, Spearman's rho non-parametric correlation test was conducted, as certain variables, such as performance and subjective reports, followed a lognormal distribution and did not conform to a normal distribution (see Figures A1 and A2 in Appendix A). The results logically revealed very high correlations between both pupils (0.96, $p < 0.01$). Significant correlations were found between pupil size (right pupil reference) and air-traffic density (0.66, $p < 0.01$), as well as between pupil size and subjective reports of mental workload (0.88, $p < 0.01$). Additionally, there was a significant correlation between pupil size and performance (0.49, $p < 0.05$). The performance measure correlated significantly with air-traffic density (0.70, $p < 0.01$) and with subjective reports (0.49, $p < 0.05$). Finally, air-traffic density and subjective reports also showed a high correlation (0.60, $p < 0.01$) (refer to Table 1).

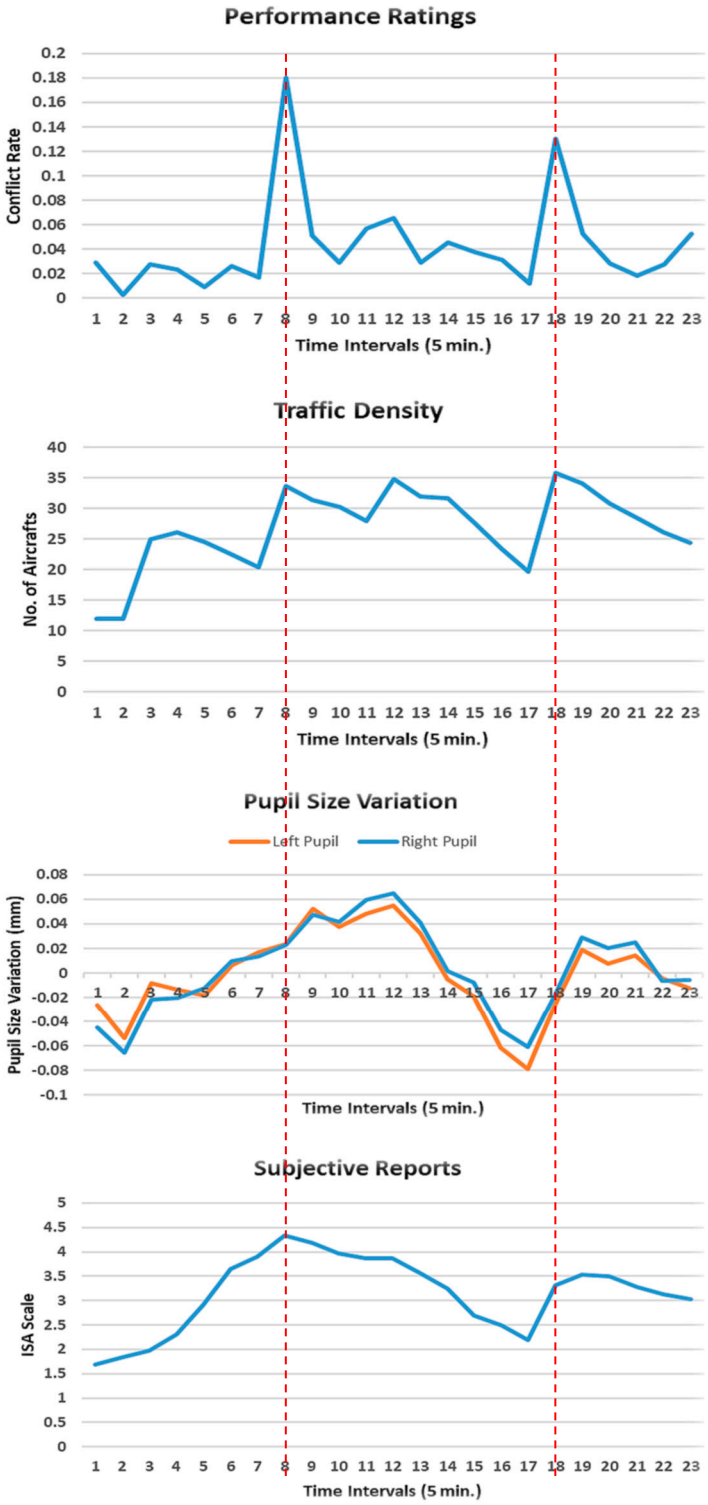


Figure 4. Performance ratings (conflict rate), air-traffic density, left and right pupil size variation and subjective mental workload reports (ISA scale) during experimental scenario development for experimental condition 1. Vertical red dotted lines indicate the position of the local maxima in conflict rate and their alignment with the remaining measures. Orange line in pupil size variation indicate left pupil, while blue line indicate right pupil.

Finally, a cross-correlation analysis was conducted to examine the latency differences between mental workload measures in reflecting poor performance peaks. This analysis was performed for pupil size (right pupil reference) and subjective reports in relation to perfor-

mance ratings, considering two intervals before and after the low-performance peaks occurred. Specifically, the low-performance peaks occurred in intervals 8 and 18, so the analysis was conducted from intervals 6 to 10 (time series 1) and from intervals 16 to 20 (time series 2).

Table 1. Mental workload measure correlation chart for experimental condition 1.

		Performance Ratings	Pupil Size (Right)	Pupil Size (Left)	Subjective Reports	Traffic Density
Performance Ratings	Spearman	1	0.489 *	0.413	0.485 *	0.695 **
	Sig. (bilateral)		0.018	0.050	0.019	0.000
	N	23	23	23	23	23
Pupil Size (Right)	Spearman	0.489 *	1	0.963 **	0.875 **	0.656 **
	Sig. (bilateral)	0.018		0.000	0.000	0.001
	N	23	23	23	23	23
Pupil Size (Left)	Spearman	0.413	0.963 **	1	0.867 **	0.602 **
	Sig. (bilateral)	0.050	0.000		0.000	0.002
	N	23	23	23	23	23
Subjective Reports	Spearman	0.485 *	0.875 **	0.867 **	1	0.596 **
	Sig. (bilateral)	0.019	0.000	0.000		0.003
	N	23	23	23	23	23
Traffic Density	Spearman	0.695 **	0.656 **	0.602 **	0.596 **	1
	Sig. (bilateral)	0.000	0.001	0.002	0.003	
	N	23	23	23	23	23

* $p < 0.05$, ** $p < 0.01$, Source: authors' elaboration.

When considering the neighboring points around the local minimum performance peaks, the highest cross-correlation values for pupil size were found at lag +1 for both time series (0.14 and 0.35, respectively). This indicates that both time series exhibit the greatest similarity when the pupil size is shifted forward one interval (5 min) relative to the local minimum performance peaks. Although this pattern was observed in both local minimum performance peaks, it is more pronounced in the second peak at interval 18 (see Figure 5).

Figure 5. Cross-correlation chart for pupil size (right pupil as reference) and performance ratings during low-performance peaks for experimental condition 1. The blue and red dotted lines show where the cross-correlation is maximized (Lag 1) and minimized (Lag −1), respectively, while the green dotted line shows the position of Lag 0.

When performing the same analysis for the cross-correlation between subjective reports and performance ratings, the best cross-correlation values were found at Lag 0 for both time series (0.28 and 0.29, respectively). These results indicate that subjective reports latency could be lower than 5 min (see Figure 6).

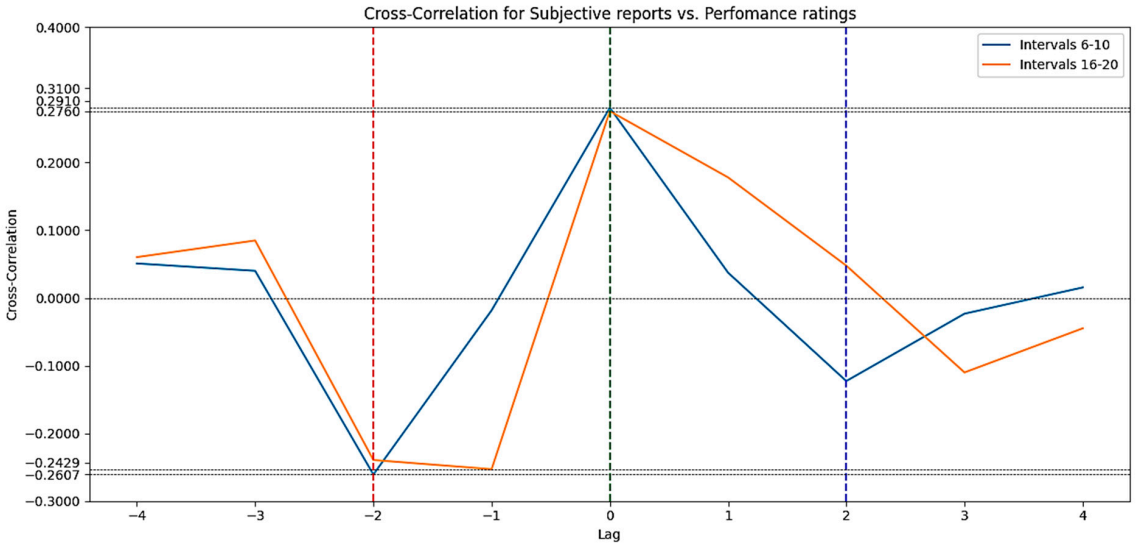


Figure 6. Cross-correlation chart for subjective reports and performance ratings during low-performance peaks for experimental condition 1. The blue and red dotted lines show where the cross-correlation is maximized (Lag 2) and minimized (Lag −2), respectively, while the green dotted line shows the position of Lag 0.

4.2. Experimental Condition 2

Experimental condition 2 was analysed similarly using a one-way within-subjects ANOVA. Figure 7 presents the charts with the results.

Similar to experimental condition 1, the performance results showed a significant effect of intervals, where $F(57,1425) = 20$, $MSe = 0.000$ and $p < 0.001$. This was primarily due to two poor performance peaks at intervals 21 and 44, while performance was nearly perfect during the other intervals. These poor performance intervals coincided with peaks in high air-traffic density, indicating higher task complexity.

Regarding pupil size, a significant effect of intervals was found for both the right and left pupils, where $F(57,1368) = 4.36$, $MSe = 0.007$ and $p < 0.001$, and $F(57,1425) = 5.28$, $MSe = 0.007$ and $p < 0.001$, respectively. Pupil size reached maximum values during the poor performance peaks, particularly from intervals 21 to 31 and 47, reflecting changes in mental workload over time.

For subjective reports of mental workload, a significant effect of intervals was found, where $F(57,1425) = 17.74$, $MSe = 0.777$ and $p < 0.001$. The graph indicates that subjective mental workload fluctuated over time, reaching maximum values during the poor performance peaks, similar to pupil size, specifically during intervals 22 and 46.

Regarding correlations between the different measures, the results revealed very high correlations between both pupils (0.94, $p < 0.01$), as expected. Significant correlations were found between pupil size (right pupil reference) and air-traffic density (0.54, $p < 0.01$), as well as between pupil size and subjective reports of mental workload (0.78, $p < 0.01$). However, the results did not show a significant correlation between pupil size and performance (0.18, $p > 0.05$). The performance measure correlated with air-traffic density (0.58, $p < 0.01$) and with subjective reports (0.39, $p < 0.01$). Finally, air-traffic density and subjective reports were highly correlated (0.75, $p < 0.01$) (refer to Table 2).

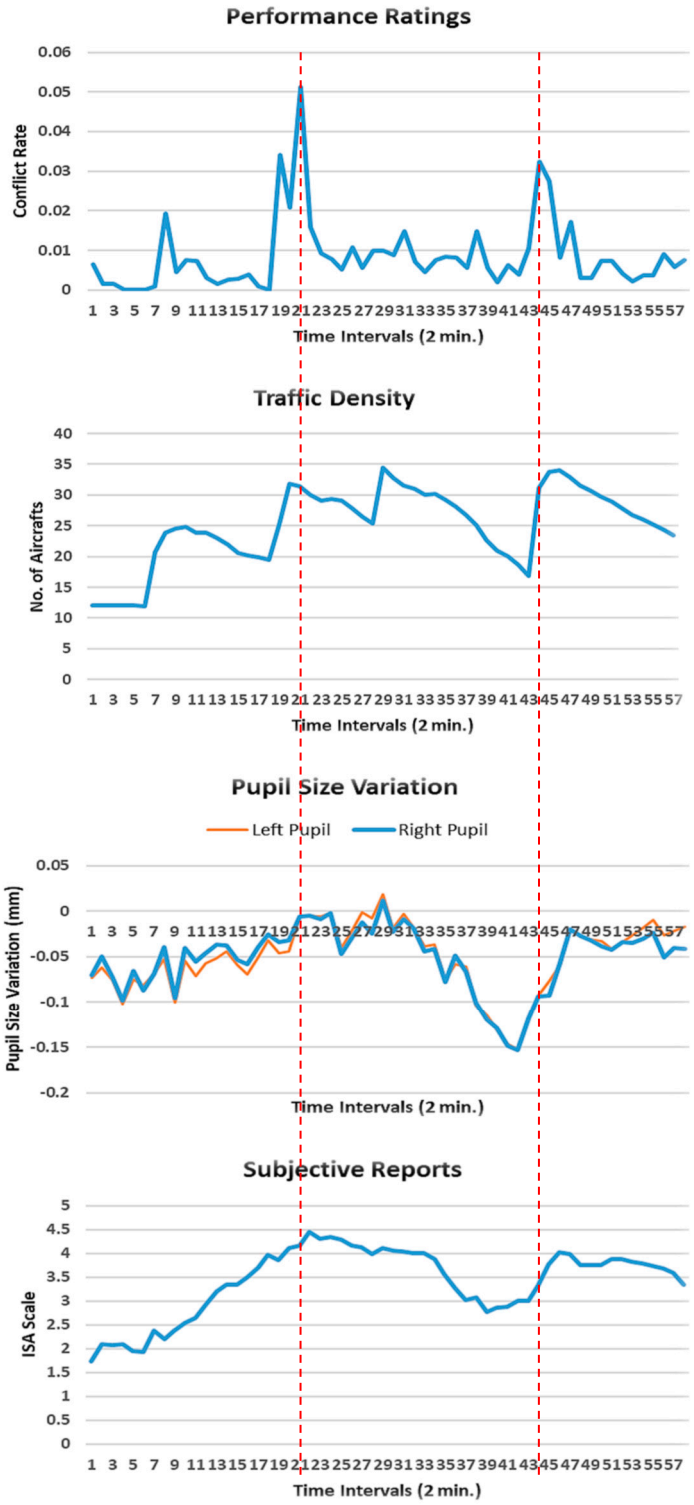


Figure 7. Performance ratings (conflict rate), air-traffic density, left and right pupil size variation and subjective mental workload reports (ISA Scale) during experimental scenario development for experimental condition 2. Vertical red dotted lines indicate the position of the local maxima in conflict rate and their alignment with the remaining measures. Orange line in pupil size variation indicate left pupil, while blue line indicate right pupil.

Again, a cross-correlation analysis was performed to examine the latency differences between mental workload measures in reflecting poor performance peaks. Specifically, the

low-performance peaks in experimental condition 2 occurred in intervals 21 and 44, so the analysis was conducted from intervals 17 to 25 and from intervals 40 to 48.

For experimental condition 2, the results were consistent with those of experimental condition 1. The highest cross-correlation values for pupil size were found at lag +2 in the first time series (0.15) and at lag +3 in the second time series (0.20), indicating a latency of 4 to 6 min for pupil size after the low-performance peak (see Figure 8).

Table 2. Mental workload measure correlation chart for experimental condition 2.

		Performance Ratings	Pupil Size (Right)	Pupil Size (Left)	Subjective Reports	Traffic Density
Performance Ratings	Spearman	1	0.177	0.218	0.390 **	0.579 **
	Sig. (bilateral)		0.184	0.100	0.003	0.000
	N	58	58	58	58	58
Pupil Size (Right)	Spearman	0.177	1	0.942 **	0.777 **	0.540 **
	Sig. (bilateral)	0.184		0.000	0.000	0.000
	N	58	58	58	58	58
Pupil Size (Left)	Spearman	0.218	0.942 **	1	0.769 **	0.534 **
	Sig. (bilateral)	0.100	0.000		0.000	0.000
	N	58	58	58	58	58
Subjective Reports	Spearman	0.390 *	0.777 **	0.769 **	1	0.745 **
	Sig. (bilateral)	0.003	0.000	0.000		0.000
	N	58	58	58	58	58
Traffic Density	Spearman	0.579 **	0.540 **	0.534 **	0.745 **	1
	Sig. (bilateral)	0.000	0.000	0.000	0.000	
	N	58	58	58	58	58

* $p < 0.05$, ** $p < 0.01$, Source: authors' elaboration.

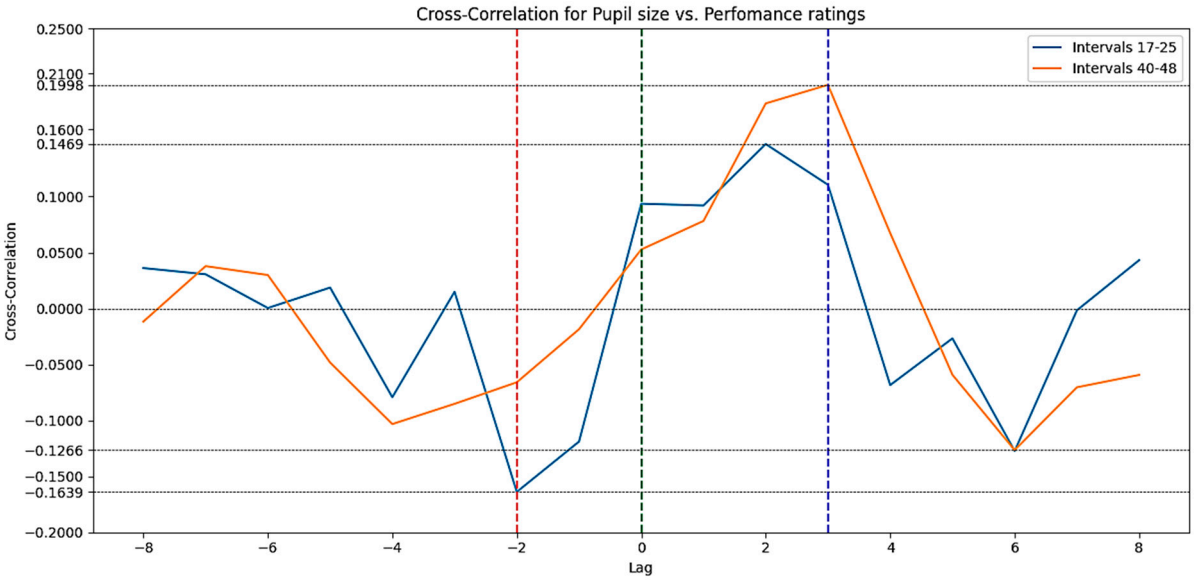


Figure 8. Cross-correlation chart for pupil size (right pupil as reference) and performance ratings during low-performance peaks for experimental condition 2. The blue and red dotted lines show where the cross-correlation is maximized (Lag 3) and minimized (Lag −2), respectively, while the green dotted line shows the position of Lag 0.

When reproducing the same analysis for the cross-correlation between subjective reports and performance ratings, the best cross-correlation values were found at lag +1 for both time series (0.19 and 0.37 respectively). This could indicate a latency in subjective reports of around 2 min in reflecting poor performance peaks, which did not appear in experimental condition 1 (see Figure 9).

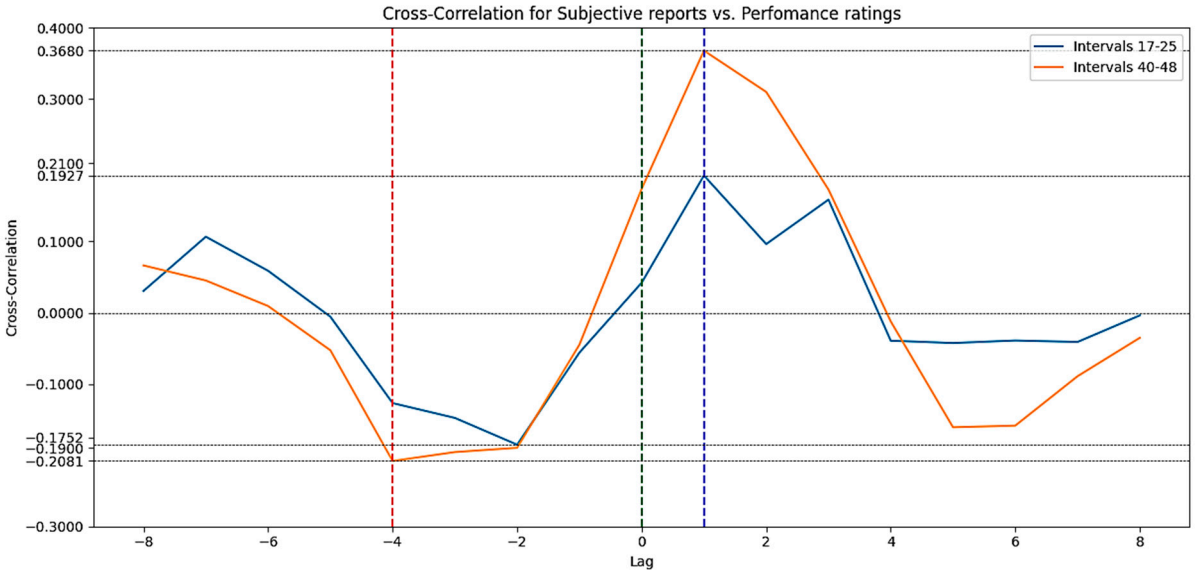


Figure 9. Cross-correlation chart for subjective reports and performance ratings during low-performance peaks for experimental condition 2. The blue and red dotted lines show where the cross-correlation is maximized (Lag 1) and minimized (Lag −4), respectively, while the green dotted line shows the position of Lag 0.

5. Discussion: Challenges and Opportunities

In this study, we explored the possibility of predicting performance and task complexity through subjective and psychophysiological mental workload measurements of participants performing an ATM task in an air-traffic control simulator. Mental workload and task complexity are indissolubly linked and the consensus in the ATM research community is that task complexity highly determines mental workload [1,2,9,10]. Hence, with this study, we wanted to explore the possibility of reversing the link between both constructs in such a way that we could derive salient characteristics determining task complexity from the observation of a raising indications of mental workload associated with a task. By monitoring subjective and psychophysiological measurements traditionally used for assessing mental workload levels, it may be possible to identify what aspects of a scenario or a task in an ATM context may foster higher task complexity and therefore performance impairment. Task complexity in the ATM domain is multifactorial. Air-traffic density certainly is one important element, but several other factors exist (standard flow interaction, potential crossings, flights in evolution, etc.) which also impact the overall task complexity, and hence would ultimately affect operators' performance, wellbeing and safety at work [9]. Understanding how, why and to what extent these different factors interfere with ATM task complexity represents one key challenge currently faced by the ATM research domain.

Both experimental conditions presented in this study are basically the same except for the granularity of the collected data. In the first experiment, the length of the intervals was set to 5 min, whereas in the second experiment, the length was set to 2 min. We manipulated our independent variable "traffic complexity" along the 120 min duration of the overall experiment by establishing four different high air-traffic density peaks.

5.1. Predicting Performance and Task Complexity through Mental Workload

The results obtained from the study showed two local performance minima at intervals 8 and 18 in the first experimental condition and at intervals 21 and 44 in the second experimental condition. These performance impairment peaks nearly coincide in both experimental conditions with subjective and psychophysiological measures of the highest observed/perceived mental workload values throughout the experimental sessions (with some lag, and this is explained in the following section), while they did not align with the variable used as a proxy for task complexity, which was air-traffic density.

Our results provided evidence on how task complexity does not only depend on air-traffic density, but on other traffic evolution and configuration variables that clearly affect task complexity. It is true that the correlation found between performance and air-traffic density is higher in both experimental conditions (0.70 and 0.58, respectively) compared to correlations found between performance and subjective reports (0.49 and 0.39, respectively) and between performance and pupil size (0.49 and 0.18, respectively). But it is also true that although air-traffic density reached four high-density peaks in both experimental conditions (at intervals 4, 8, 12 and 18 in the first experimental condition and at intervals 10, 20, 29 and 46 in the second); we can see in the results that we only obtained two real high-complexity peaks that would be shown as local minimum performance peaks at intervals 8 and 18 in the first experimental condition and at intervals 21 and 44 in the second. This evidence was not surprising, as it aligns with literature research [1–3,9,72–74] and shows that we cannot just trust air-traffic density for predicting traffic complexity and operators' bad performance peaks, but rather we must understand task complexity in the ATM domain as a combination of multiple factors interacting with each other. Understanding how these multiple factors affect task complexity is essential and will be a very relevant issue in future ATM research.

An important finding of this study is that we found associations between mental workload measurements and task-load levels; both mental workload indexes, subjective and psychophysiological, showed correlations with external task load. Both indexes increased as task complexity increased (represented by performance impairment peaks) and both decreased as task complexity decreased as well. Hence, the key finding of this study is that mental workload measurements predicted performance impairment peaks (and air-traffic high-complexity peaks) successfully, whereas air-traffic density did not, even though correlations between performance and air-traffic density were higher than between performance and subjective/psychophysiological measures. The main implication of this finding is, on the one hand, that we should rely more on mental workload measurements for predicting task complexity peaks in the ATM domain rather than simply trusting air-traffic density; and, on the other hand, that we may use mental workload measurements in further research for analyzing and determining which, how and to what extent the different air-traffic factors affect task complexity in ATM.

5.2. Latency Differences between Measurements

Another significant finding of the study is that we observed latency differences between mental workload measurements in reflecting task-load variations, particularly during low-performance peaks. The cross-correlation analysis between mental workload measures and performance ratings revealed that the subjective measurement showed a lower latency than the psychophysiological measurement in reflecting low-performance peaks. In experimental condition 1, the best cross-correlation values for the subjective reports were found at Lag 0 for both low-performance peaks. Similarly, in experimental condition 2, the best cross-correlation values for the subjective ratings were found at lag +1 for both low-performance peaks. These results suggest a latency in subjective reports of around 0 to 2 min in reflecting poor performance peaks. Conversely, considering pupil size, the best cross-correlation values were found at lag +1 in experimental condition 1 for both low-performance peaks and at lag +2 and +3 in experimental condition 2 for each

low-performance peak, respectively. These findings indicate that pupil size may have a latency of about 4 to 6 min in reflecting low-performance peaks.

Overall, subjective measurements seem to reflect poor performance peaks significantly earlier than pupil size measurements. Therefore, although we found associations between task demand and mental workload measurements, we must be cautious and consider that differences in latency exist between measures when reflecting a task-load variation and the reason provided by Hancock for this is that each different method possesses its own differential timescale; one mental workload measure can reflect changes almost immediately, while some others could show a higher latency when reflecting a mental workload change over time. The main implication of this finding is that we must be cautious when using mental workload measures for predicting performance and task complexity in such a way that we must consider the latency of each measure in reflecting task demand changes. We should also highlight the negative implications of latency differences among measures in real-world ATM operations. For instance, relying on a single mental workload measure with high latency in reflecting workload changes could be dangerous for advising ATCos about their current mental workload status, especially if they are used to being monitored by measures with lower latency. An overload situation could be reached before the index detects it, leading to risky scenarios; ATCOs' decision-making would be negatively affected under overload, and performance drops could occur, both increasing the likelihood of incidents and accidents. To avoid this, a thorough understanding of the specific latencies of different mental workload measures is essential. Dissociations between measurements due to latency and other factors are currently a key research topic in the human factors and ergonomics field [75–77].

Finally, we noticed an unexpected finding in our results, which was observed in both experimental conditions. After looking at our data, we noticed that there exists a task learning effect that affects participants' mental workload; during the first high-demand peak, the participants showed a higher mental workload peak than during the second peak of high demand and this was reflected in both subjective and psychophysiological measures. This was due to the fact that at the beginning, the participants were new to the experimental scenario, but after practicing with it for a certain amount of time, they managed to develop new cognitive strategies to cope with it; in other words, there was a shift to a more automated task performance, which reduced the experienced mental workload. These results could be explained by the "cognitive load theory" [78].

5.3. Limitations, Final Conclusions and Further Research

However, several limitations must be considered. First, this study was conducted considering only three different mental workload measurements, i.e., the ISA scale as our subjective perceived mental workload measurement, the conflict rate as our performance measurement and the pupil size as our psychophysiological mental workload measurement. There are dozens of mental workload measurements that could have been used for this research, each with their own response latency and behavior. It would be very interesting to analyze how these different measurements react to changes in task demand and assess their appropriateness and validity for predicting performance and task complexity. Examples of psychophysiological measures that could be used include electroencephalogram (EEG) measures, heart-rate variability, electrodermal activity, ocular parameters (e.g., blink rate or gaze mapping), and voice parameters (e.g., fundamental frequency of the voice; speech rate). This should be urgently addressed in further research to determine which mental workload measurements best suit the scope of our study. For instance, it would be valuable to analyze which specific psychophysiological measure, from a wide variety, (1) correlates best with performance and task complexity and (2) shows the smallest latency in reflecting MW changes, given the evident benefits of this. Second, another issue considered was the granularity of collected data. The authors therefore designed two identical experimental conditions, one where the ISA values were collected every 5 min and one in which the values were collected every 2 min. The underlying rationale was to determine if greater

granularity could offer improved insights into the correlation among the three different primary measures of mental workload considered. However, shortening the data collection interval did not provide the expected benefit as the Spearman correlation values reported for the observed variables (subjective reports, pupil size, traffic density) in respect to the target variable (performance ratings) were higher for the 5 min interval experimental setting than for the 2 min one. This is justified by the fact that the intrusiveness implied with the shorter data collection intervals might have impacted the quality of the collected data affecting the participant mental workload and mental fatigue within the task development. Finally, this research was conducted with a sample of students under very controlled laboratory settings. Since our aim was to study basic psychological processes, a sample of students in a simulated context was adequate for this purpose. However, it is also true that these results cannot be fully extrapolated to real ATCOs in actual working environments. Therefore, to improve the external validity of our findings, this study should be replicated with a more representative sample and in more applied contexts. Specifically, it would be ideal to repeat this study with real ATCOs performing their ATM tasks in real working environments, as such results would have much higher external and ecological validity.

With this study, we wanted to shed some light on the possibility of using mental workload measurements as predictors of task performance and complexity, in such a way that we could ultimately use them to carry out research on which particular ATM factors have a greater effect on the dynamics of ATM task complexity. Our results revealed that mental workload measurements better predict minimum performance and high task complexity peaks than other supposed well-established factors such as the air-traffic density, and so they could be used for carrying out research on understanding how the different ATM factors affect task complexity. It is widely known in the ATM research sphere that task complexity does not only depend on air-traffic density, but on so many other factors, which affect the configuration and evolution of the flights in the radar screen. Understanding the role and the weight of these factors in the overall task complexity confronted by ATCos constitutes one key subject that is being researched both now and in the future. We are convinced that mental workload measurements can be used as valid and reliable tools for this purpose; however, further research must be carried out to continue improving our knowledge about how we can better use mental workload different measurements for predicting performance and task complexity and, hence, for developing a model for predicting ATM complexity built on the basis of the different contribution of ATM factors to the overall complexity.

Author Contributions: Conceptualization, E.M.-d.-E., M.C.L. and J.J.C.; methodology, E.M.-d.-E., M.C.L. and J.J.C.; validation, E.M.-d.-E., M.C.L. and J.J.C.; formal analysis, E.M.-d.-E., M.C.L. and J.J.C.; investigation, E.M.-d.-E., M.C.L. and J.J.C.; resources, E.M.-d.-E., M.C.L. and J.J.C.; writing—original draft preparation, E.M.-d.-E., M.C.L. and J.J.C.; writing—review and editing, E.M.-d.-E., M.C.L. and J.J.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research conducted in this publication was funded by the Irish Research Council under grant number [EPSPD/2022/151].

Data Availability Statement: We made the data shown in this study freely and openly accessible at https://doi.org/10.6084/m9.figshare.17433440.

Acknowledgments: We would like to express our sincere gratitude to Houda Briwa and Andrés Alonso Perez for their invaluable contributions to this manuscript. Their efforts were instrumental in readapting the methodology, conducting the formal data analysis, and assisting with the writing, review, and editing process. Specifically, their expertise was critical in fulfilling the reviewers' suggestions, which included the addition of a new statistical analysis—cross-correlation analysis—to meet the requirements of one of the manuscript reviewers. Their dedication and expertise significantly improved the quality and rigor of this work.

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

Appendix A

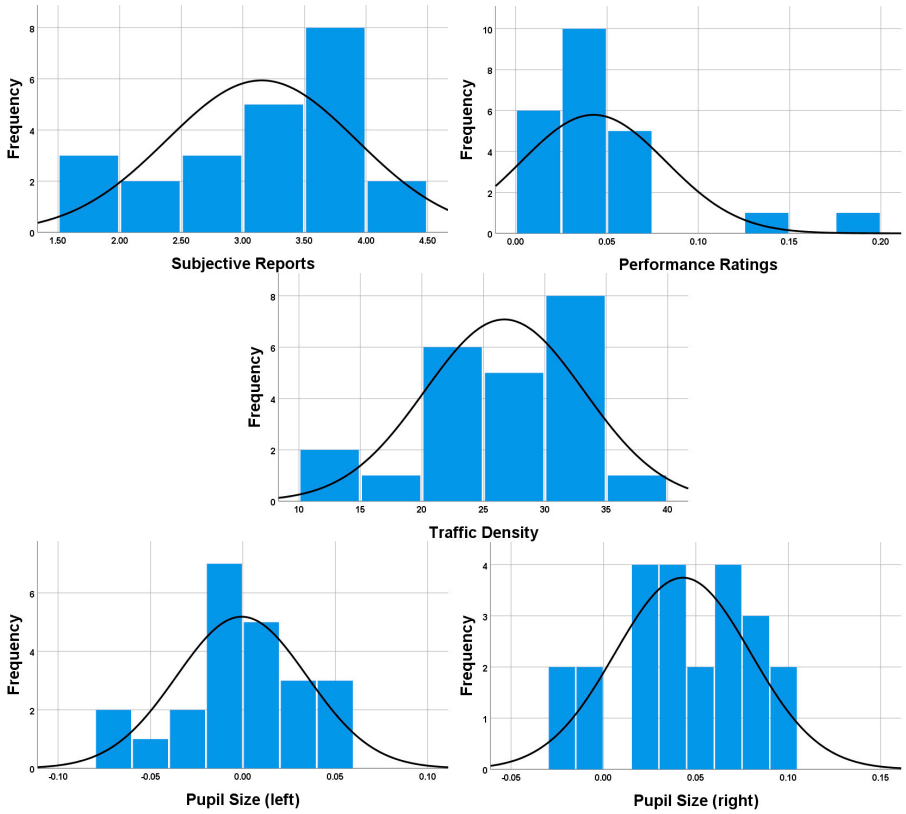


Figure A1. Histograms for subjective mental workload reports (ISA scale), performance ratings (conflict rate), air-traffic density and left and right pupil size for experimental condition 1.

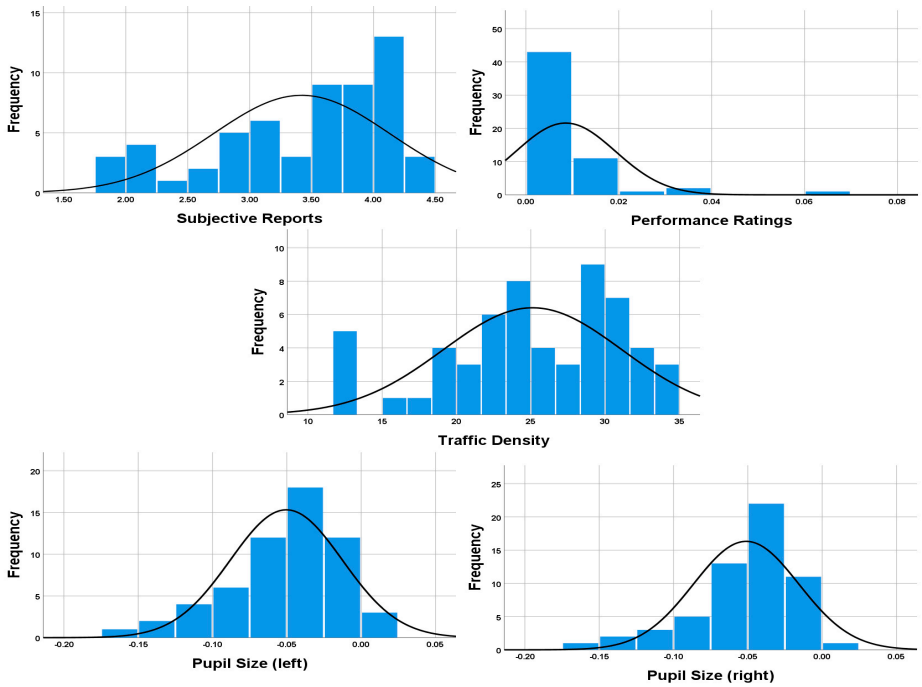


Figure A2. Histograms for subjective mental workload reports (ISA scale), performance ratings (conflict rate), air-traffic density and left and right pupil size for experimental condition 2.

References

1. Hilburn, B. Cognitive complexity in air traffic control: A literature review. *EEC Note* **2004**, *4*, 1–80.
2. Histon, J.M.; Hansman, R.J. Mitigating Complexity in Air Traffic Control: The Role of Structure-Based Abstractions. Ph.D. Thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA, 2008.
3. Cañas, J.J.; Ferreira, P.; de Frutos, P.L.; Puntero, E.; López, E.; Gómez-Comendador, F.; de Crescenzio, F.; Lucchi, F.; Netjasov, F.; Mirkovic, B. Mental workload in the explanation of automation effects on ATC performance. In *International Symposium on Human Mental Workload: Models and Applications*; Springer: Cham, Switzerland, 2018; pp. 202–221.
4. Edwards, T.; Homola, J.; Mercer, J.; Claudatos, L. Multifactor interactions and the air traffic controller: The interaction of situation awareness and workload in association with automation. *Cogn. Technol. Work.* **2017**, *19*, 687–698. [CrossRef]
5. Edwards, T.; Gabets, C.; Mercer, J.; Bienert, N. Task Demand Variation in Air Traffic Control: Implications for Workload, Fatigue, and Performance. In *Advances in Human Aspects of Transportation*; Stanton, N.A., Landry, S., Di Bucchianico, G., Vallicelli, A., Eds.; Springer: Cham, Switzerland, 2017; Volume 484, pp. 91–102. ISBN 9783319416816.
6. Alaydi, B.; Ng, S.-I. Mitigating the Negative Effect of Air Traffic Controller Mental Workload on Job Performance: The Role of Mindfulness and Social Work Support. *Safety* **2024**, *10*, 20. [CrossRef]
7. Balta, E.; Psarrakis, A.; Vatakis, A. The effects of increased mental workload of air traffic controllers on time perception: Behavioral and physiological evidence. *Appl. Ergon.* **2024**, *115*, 104162. [CrossRef] [PubMed]
8. Amalberti, R. *La Conduite de Systèmes à Risques*; Presses Universitaires de France: Paris, France, 1996.
9. de Frutos, P.L.; Rodríguez, R.R.; Zhang, D.Z.; Zheng, S.; Cañas, J.J.; Muñoz-de-Escalona, E. COMETA: An air traffic controller's mental workload model for calculating and predicting demand and capacity balancing. In *International Symposium on Human Mental Workload: Models and Applications*; Springer: Cham, Switzerland, 2019; pp. 85–104.
10. Netjasov, F.; Janić, M.; Tošić, V. Developing a generic metric of terminal airspace traffic complexity. *Transportmetrica* **2011**, *7*, 369–394. [CrossRef]
11. Dawson, D.; Noy, Y.I.; Härmä, M.; Åkerstedt, T.; Belenky, G. Modelling fatigue and the use of fatigue models in work settings. *Accid. Anal. Prev.* **2011**, *43*, 549–564. [CrossRef]
12. Sarsangi, V.; Salehiniya, H.; Hannani, M.; Marzaleh, M.A.; Abadi, Y.S.; Honarjoo, F.; Derakhshanjazari, M. Assessment of workload effect on nursing occupational accidents in hospitals of Kashan, Iran. *Biomed. Res. Ther.* **2017**, *4*, 1527–1540. [CrossRef]
13. Smith, A.P.; Smith, H.N. Workload, fatigue and performance in the rail industry. In *International Symposium on Human Mental Workload: Models and Applications*; Springer: Cham, Switzerland, 2017; pp. 251–263. [CrossRef]
14. Berberian, B.; Somon, B.; Sahaï, A.; Gouraud, J. The out-of-the-loop Brain: A neuroergonomic approach of the human automation interaction. *Annu. Rev. Control* **2017**, *44*, 303–315. [CrossRef]
15. Di Flumeri, G.; De Crescenzio, F.; Berberian, B.; Ohneiser, O.; Kramer, J.; Aricò, P.; Borghini, G.; Babiloni, F.; Bagassi, S.; Piastra, S. Brain–computer interface-based adaptive automation to prevent out-of-the-loop phenomenon in air traffic controllers dealing with highly automated systems. *Front. Hum. Neurosci.* **2019**, *13*, 296. [CrossRef] [PubMed]
16. Cañas, J.J.; Ferreira, P.N.P.; Puntero, E.; López, P.; López, E.; Gomez-Comendador, V.F. An air traffic controller psychological model with automation. In Proceedings of the 7th EASN International Conference: "Innovation in European Aeronautics Research", Warsaw, Poland, 26–29 September 2016.
17. Majumdar, A.; Ochieng, W. Factors affecting air traffic controller workload: Multivariate analysis based on simulation modeling of controller workload. *Transp. Res. Rec.* **2002**, *1788*, 58–69. [CrossRef]
18. Wu, C.; Liu, Y. Queuing network modeling of driver workload and performance. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 528–537.
19. Sozou, P.D.; Lane, P.C.; Addis, M.; Gobet, F. Computational scientific discovery. In *Handbook of Model-Based Science*; Magnani, L., Bertolotti, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 719–734.
20. Longo, L.; Leva, M.C. (Eds.) *Human Mental Workload: Models and Applications: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, 20–21 September 2018, Revised Selected Papers*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1012.
21. Longo, L.; Leva, M.C. (Eds.) *Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, 14–15 November 2019, Proceedings*; Springer Nature: Berlin/Heidelberg, Germany, 2019; Volume 1107.
22. Longo, L.; Leva, M.C. (Eds.) *Human Mental Workload: Models and Applications: 4th International Symposium, H-WORKLOAD 2020, Granada, Spain, 3–5 December 2020: Proceedings*; Springer Nature: Berlin/Heidelberg, Germany, 2020; Volume 1318.
23. Moustafa, K.; Luz, S.; Longo, L. Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In Proceedings of the International Symposium on Human Mental Workload: Models and Applications, Dublin, Ireland, 28–30 June 2017; Springer: Cham, Switzerland, 2017; pp. 30–50.
24. Rizzo, L.; Longo, L. Representing and inferring mental workload via defeasible reasoning: A comparison with the NASA Task Load Index and the Workload Profile. In Proceedings of the 1st Workshop on Advances in Argumentation in Artificial Intelligence, Bari, Italy, 15–16 November 2017.
25. Rizzo, L.; Dondio, P.; Delany, S.J.; Longo, L. Modeling mental workload via rule-based expert system: A comparison with NASA-TLX and workload profile. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Thessaloniki, Greece, 16–18 September 2016; Springer: Cham, Switzerland, 2016; pp. 215–229.
26. Crevits, I.; Debernard, S.; Denecker, P. Model building for air-traffic controllers' workload regulation. *Eur. J. Oper. Res.* **2002**, *136*, 324–332. [CrossRef]

27. Moray, N. *Mental workload: Its Theory and Measurement*; Plenum Press: New York, NY, USA, 1979.
28. Wickens, C.D. Mental workload: Assessment, prediction and consequences. In Proceedings of the International Symposium on Human Mental Workload: Models and Applications, Dublin, Ireland, 28–30 June 2017; Springer: Cham, Switzerland, 2017; pp. 18–29.
29. Murai, K.; Hayashi, Y.; Okazaki, T.; Stone, L.C. Evaluation of ship navigator's mental workload using nasal temperature and heart rate variability. In Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12–15 October 2008; IEEE: New York, NY, USA, 2008; pp. 1528–1533.
30. De Alwis Edirisinghe, V. Estimating Mental Workload of University Students Using Eye Parameters. Master's Thesis, NTNU, Trondheim, Norway, 2017.
31. Hancock, P.A. Whither workload? mapping a path for its future development. In Proceedings of the International Symposium on Human Mental Workload: Models and Applications, Dublin, Ireland, 28–30 June 2017; Springer: Cham, Switzerland, 2017; pp. 3–17.
32. Yeh, Y.H.; Wickens, C.D. *The Dissociation of Subjective Measures of Mental Workload and Performance*; National Aeronautics and Space Administration, Ames Research Center: Silicon Valley, CA, USA, 1984.
33. Casper, P.A. Dissociations among Measures of Mental Workload: Effects of Experimenter-Induced Inadequacy. Ph.D. Thesis, Purdue University, West Lafayette, IN, USA, 1988.
34. Fothergill, S.; Loft, S.; Neal, A. ATC-labAdvanced: An air traffic control simulator with realism and control. *Behav Res. Methods* **2009**, *41*, 118–127. [CrossRef] [PubMed]
35. Loft, S.; Finnerty, D.; Remington, R.W. Using spatial context to support prospective memory in simulated air traffic control. *Hum. Factors* **2011**, *53*, 662–671. [CrossRef] [PubMed]
36. Gee, D.G.; Bath, K.G.; Johnson, C.M.; Meyer, H.C.; Murty, V.P.; van den Bos, W.; Hartley, C.A. Neurocognitive development of motivated behavior: Dynamic changes across childhood and adolescence. *J. Neurosci.* **2018**, *38*, 9433–9445. [CrossRef]
37. Yeo, G.B.; Frederiks, E.R.; Kiewitz, C.; Neal, A. A dynamic, self-regulatory model of affect and performance: Interactions between states, traits and task demands. *Motiv. Emot.* **2014**, *38*, 429–443. [CrossRef]
38. Wilson, M.D.; Strickland, L.; Farrell, S.; Visser, T.A.; Loft, S. Prospective memory performance in simulated air traffic control: Robust to interruptions but impaired by retention interval. *Hum. Factors* **2020**, *62*, 1249–1264. [CrossRef]
39. Marchitto, M.; Di Stasi, L.L.; Cañas, J.J. Ocular movements under taskload manipulations: In-fluence of geometry on saccades in air traffic control simulated tasks. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2012**, *22*, 407–419. [CrossRef]
40. Soeadyfa Fridyatama, D.A.; Suparji, S.; Sumbawati, M.S. Developing Air Traffic Control Simulator for Laboratory. *TEM J.* **2023**, *12*, 1462–1474. [CrossRef]
41. Brennan, S.D. *An Experimental Report on Rating Scale Descriptor Sets for the Instantaneous Self Assessment (ISA) Recorder*; DRA Technical Memorandum (CAD5); DRA Maritime Command and Control Division: Portsmouth, UK, 1992; p. 92017.
42. Jordan, C.S. *Experimental Study of the Effect of an Instantaneous Self Assessment Workload Recorder on Task Performance*; DRA Technical Memorandum (CAD5); DRA Maritime Command Control Division: Portsmouth, UK, 1992; p. 92011.
43. Prandini, M.; Piroddi, L.; Puechmorel, S.; Brázdilová, S.L. Toward air traffic complexity assessment in new generation air traffic management systems. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 809–818. [CrossRef]
44. Kartali, A.; Janković, M.M.; Gligorijević, I.; Mijović, P.; Mijović, B.; Leva, M.C. Real-time mental workload estimation using eeg. In *Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, 14–15 November 2019, Proceedings 3*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 20–34.
45. Socha, V.; Hanáková, L.; Valenta, V.; Socha, L.; Ábela, R.; Kušmírek, S.; Pilmannová, T.; Tecl, J. Workload assessment of air traffic controllers. *Transp. Res. Procedia* **2020**, *51*, 243–251. [CrossRef]
46. Díaz Robredo, L.A.; Robles Sánchez, J.I. La actividad electrodérmica de la piel como indicador de activación psicofisiológica en pilotos de caza españoles: Un estudio preliminar. *Sanid. Mil.* **2018**, *74*, 7–12.
47. Lenné, M.G.; Jacobs, E.E. Predicting drowsiness-related driving events: A review of recent research methods and future opportunities. *Theor. Issues Ergon. Sci.* **2016**, *17*, 533–553. [CrossRef]
48. Lee, C.W.; Cuijpers, P. A meta-analysis of the contribution of eye movements in processing emotional memories. *J. Behav. Ther. Exp. Psychiatry* **2013**, *44*, 231–239. [CrossRef]
49. Tsai, Y.F.; Viirre, E.; Strychacz, C.; Chase, B.; Jung, T.P. Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviat. Space Environ. Med.* **2007**, *78*, B176–B185. [PubMed]
50. Zheng, B.; Jiang, X.; Tien, G.; Meneghetti, A.; Panton ON, M.; Atkins, M.S. Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surg. Endosc.* **2012**, *26*, 2746–2750. [CrossRef]
51. Matthews, G.; Middleton, W.; Gilmartin, B.Y.; Bullimore, M.A. Pupillary diameter and cognitive and cognitive load. *J. Psychophysiol.* **1991**, *5*, 265–271.
52. Backs RW, Y.; Walrath, L.C. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Appl. Erg.* **1992**, *23*, 243–254. [CrossRef]
53. Hyönä, J.; Tommola, J.; Alaja, A. Pupil dilation as a measure of processing load in simultaneous interpreting and other language tasks. *Q. J. Exp. Psychol.* **1995**, *48*, 598–612. [CrossRef]
54. Granholm, E.; Asarnow, R.F.; Sarkin, A.J.; Dykes, K.L. Pupillary responses index cognitive resource limitations. *Psychophysiology* **1996**, *33*, 457–461. [CrossRef]

55. Iqbal, S.T.; Zheng, X.S.; Bailey, B.P. Task evoked pupillary response to mental workload in human-computer interaction. In Proceedings of the ACM Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; ACM: New York, NY, USA, 2004; pp. 1477–1480.
56. Verney, S.P.; Granholm, E.; Marshall, S.P. Pupillary responses on the visual backward masking task reflect general cognitive ability. *Int. J. Psychophysiol* **2004**, *52*, 23–36. [CrossRef] [PubMed]
57. Porter, G.; Troscianko, T.; Gilchrist, I.D. Effort during visual search and counting: Insights from pupillometry. *Q. J. Exp. Psychol* **2007**, *60*, 211–229. [CrossRef]
58. Privitera, C.M.; Renninger, L.W.; Carney, T.; Klein, S.; Aguilar, M. Pupil dilation during visual target detection. *J. Vis.* **2010**, *10*, 3. [CrossRef]
59. Piquado, T.; Isaacowitz, D.; Wingfield, A. Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology* **2010**, *47*, 560–569. [CrossRef] [PubMed]
60. Reiner, M.; Gelfeld, T.M. Estimating mental workload through event-related fluctuations of pupil area during a task in a virtual world. *Int. J. Psychophysiol.* **2014**, *93*, 38–44. [CrossRef] [PubMed]
61. Hess, E.H.; Polt, J.M. Pupil size in relation to mental activity during simple problem-solving. *Science* **1964**, *143*, 1190–1192. [CrossRef] [PubMed]
62. Elshtain, L.; Schaefer, T. Effects of storage load and word frequency on pupillary responses during short-term memory. *Psychon. Sci.* **1968**, *12*, 143–144. [CrossRef]
63. Peavler, W.S. Pupil size, information overload, and performance differences. *Psychophysiology* **1974**, *11*, 559–566. [CrossRef]
64. Bradshaw, J. Pupil size as a measure of arousal during information sing. *Nature* **1967**, *216*, 515–516. [CrossRef]
65. Payne, D.T.; Parry, M.E.; Harasymiw, S.J. Percentage of pupillary dilation as a measure of item difficulty. *Percept. Psychophys.* **1968**, *4*, 139–143. [CrossRef]
66. Schaefer, T.; Ferguson, J.B.; Klein, J.A.; Rawson, E.B. Pupillary responses during mental activities. *Psychon. Sci.* **1968**, *12*, 137–138. [CrossRef]
67. Bradshaw, J. Pupil size and problem solving. *Q. J. Exp. Psychol.* **1968**, *20*, 116–122. [CrossRef]
68. Kahneman, D. Attention and effort. In *Englewood Cliffs*; Prentice-Hall: Upper Saddle River, NJ, USA, 1973; Volume 1063.
69. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **1982**, *91*, 276–292. [CrossRef]
70. Sirois, S.; Brisson, J. Pupillometry. *Wiley Interdiscip. Rev. Cogn. Sci.* **2014**, *5*, 679–692. [CrossRef]
71. Mathôt, S.; Fabius, J.; Van Heusden, E.; Van der Stigchel, S. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav. Res. Methods* **2018**, *50*, 94–106. [CrossRef] [PubMed]
72. Mogford, R.H.; Guttman, J.A.; Morrow, S.L.; Kopardekar, P. *The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature*; Cta Inc.: Mckee City, NJ, USA, 1995.
73. Athènes, S.; Averty, P.; Puechmorel, S.; Delahaye, D.; Collet, C. ATC complexity and controller workload: Trying to bridge the gap. In Proceedings of the International Conference on HCI in Aeronautics; AAAI Press: Cambridge, MA, USA, 2002; pp. 56–60.
74. Djokic, J.; Lorenz, B.; Fricke, H. Air traffic control complexity as workload driver. *Transp. Res. Part C Emerg. Technol.* **2010**, *18*, 930–936. [CrossRef]
75. Muñoz-de-Escalona, E.; Cañas, J.J.; van Nes, J. Task Demand Transition Rates of Change Effects on Mental Workload Measures Divergence. In Proceedings of the International Symposium on Human Mental Workload: Models and Applications, Rome, Italy, 14–15 November 2019; Springer: Cham, Switzerland, 2019; pp. 48–65.
76. Muñoz-de-Escalona, E.; Cañas, J.J.; Leva, C.; Longo, L. Task Demand Transition Peak Point Effects on Mental Workload Measures Divergence. In Proceedings of the International Symposium on Human Mental Workload: Models and Applications, Granada, Spain, 3–5 December 2020; Springer: Cham, Switzerland, 2020; pp. 207–226.
77. Hancock, P.A.; Matthews, G. Workload and performance: Associations, insensitivities, and dissociations. *Hum. Factors* **2019**, *61*, 374–392. [CrossRef]
78. Sweller, J. Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* **1994**, *4*, 295–312. [CrossRef]