Journal of Air Transport Management 108 (2023) 102378

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman

# Experimental validation of COMETA model of mental workload in air traffic control☆

Jorge Ibáñez-Gijón [a], David Travieso [a,*], José A. Navia [a], Aitor Montes [a], David M. Jacobs [a], Patricia L. Frutos [b]

[a] Dpto Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, Madrid, Spain
[b] CRIDA A.I.E. ATM R&D + Innovation Reference Centre, Madrid, Spain

ARTICLE INFO

*Keywords:*
Air traffic control
Mental workload model
Task complexity
Transport ergonomics
System performance

ABSTRACT

The sustained increase in air traffic during the last decades represents a challenge to the air traffic management system in general. Thus, it is of utmost importance to develop strategies that can safely increase air traffic controller's handling capacity without increasing task related strain. This research proposes and validates a predictive model of air traffic controller's mental workload. Our model is based on COMETA, a model that considers the effect of the most relevant air traffic events in the cognitive complexity of the task. In the version of COMETA used in this study we include the online effects of the controllers' actions on the state of the airspace. To validate the model, a laboratory experiment was conducted using a simulator to precisely control the task workload factors. We used traffic density and airspace complexity as experimental factors because they are the most commonly acknowledged sources of mental workload in air traffic control literature. The measured dependent variables were selected because they have been found to correlate with mental workload in ATC tasks, namely, ISA and NASA indexes, electrodermal activity, heart rate, and different performance measures. The results demonstrate that our model can successfully predict air traffic controllers' mental workload across a wide range of task workload conditions. In addition, our results provide a clear portrait of the complex interactions between the different sources of task workload and their effects on mental workload. In the conclusion we consider the limitations and opportunities for the application of this model to improve policies.

## 1. Introduction

Air traffic controllers (ATCOs) perform a highly demanding task with a crucial role on the functioning of the air traffic management -ATM- system (Durso and Manning, 2008; Hilburn, 2004; Hopkin, 2017). In addition, the sustained increase in air traffic during the last decades poses a challenge to the ATM system (ICAO International Civil Aviation Organization, 2007; Matsumoto, 2007). Thus, increasing the aircraft handling rate has become a strategic milestone for current ATM systems, to which the ATCOs workload is the key limiting factor (Djokic et al., 2010; EUROCONTROL, 2004). There are three fundamental approaches to achieve this goal: increase the pool of ATCOs, optimize workload distribution among the members of the pool, and increase the capacity of each ATCO to handle traffic.

Increasing ATCOs handling capacity must be performed safely. For example, they may use assistive technologies and new ATM systems to relief the controller from part of the burden posed by the task. It is also possible to increase the ATCO handling capacity by optimizing the spatiotemporal workload distribution. These solutions have in common that they would benefit from a more precise understanding of mental workload in ATCOs (Aricò et al., 2017; Loft et al., 2007; Majumdar et al., 2004; Metzger and Parasuraman, 2017; Mitchell, 2000; Mohammed and El Bekkaye, 2021).

Thus, the objective of this study is to propose and validate a model of mental workload (MWL) that encompasses the most relevant sources of task workload (TWL) and the cognitive strategies that the ATCOs can use to deal with them. The proposed model considers the real time effects of the ATCO decisions on the dynamic evolution of the airspace. Such model can be helpful to predict and avoid the most complex situations in ATC tasks. To validate the model, we have performed an experimental

☆ This research was supported by funds from the Agreement UAM-ENAIRE-CRIDA, for the development of R + D activities on Human Factors in Air Traffic Management (BOE-A-2019-11691).

* Corresponding author. Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, 28049, Madrid, Spain.
*E-mail address:* david.travieso@uam.es (D. Travieso).

https://doi.org/10.1016/j.jairtraman.2023.102378
Received 30 July 2022; Received in revised form 30 December 2022; Accepted 2 February 2023
Available online 11 February 2023

study using a simulated ATC task that allowed us to precisely manipulate the main sources of task workload. We next present an overview of the research on MWL in ATC.

### 1.1. Workload in ATC

Workload is a relatively recent concept that was introduced to explain the inability of human operators to cope with the requirements of a task (Gopher and Braune, 1984). It refers to a complex relational process that emerges from the multiscale interactions between a specific agent and the task being performed (Cain, 2007; Huey and Wickens, 1993; Ibáñez-Gijón et al., 2017; Pagnotta et al., 2021). It is often considered as the effort that an agent has to exert in order to perform a task (Hart, 2006). In the study of physical workload, it is regarded as composed of two elements: the stress or TWL and the strain. Stress is conceived as the task demands independent of the agent, whereas strain is used to denote their impact on the agent. In this framework, MWL would refer to strain caused by cognitive tasks. However, applying this clear cut distinction is not so simple for tasks that involve more than bare physical effort (Wickens, 2008). Task demands cannot be meaningfully expressed independently of the specific agent, nor the agent skills can be understood as general purpose functions detached from the actual tasks in which they are used (Cain, 2007; Ibáñez-Gijón et al., 2017).

The distinction between task and mental workload is nevertheless useful, as demonstrated by its daily use in standardization regulations. TWL is comparatively easier to measure because it is determined by the objective dimensions of the task. On the other hand, MWL is not a measurable quantity because it is the mental effort required to perform a certain task. One can only infer MWL by its effects on other processes, by asking the agent, or by observing the outcomes of the agent-task interaction. In addition, MWL is strongly dependent on previous experiences and can be dramatically affected by switches of the strategies used by the ATCOs or their situational awareness (Edwards et al., 2017; Endsley, 1995). Therefore, the scientific study of MWL requires theoretical models to estimate MWL from detailed descriptions of TWL and expert knowledge of the strategies used to solve the task (Fürstenau and Radüntz, 2022; Mitchell, 2000). To accommodate the relational nature of workload, such models should attempt to be as specific as possible to the concrete agent-task system in interaction that brings about MWL, rather than detached universal problem solvers (Bingham, 1988). To anticipate, in this research we propose and validate a model of MWL based on integration of expert knowledge about ATC TWL and the cognitive processes that the ATCOs can use to reduce MWL.

### 1.2. COMETA model of ATCO MWL

COMETA (COgnitive ModEl for aTco workload Assessment) is a model developed by CRIDA to predict the ATCO MWL (Frutos et al., 2019). Its starting point is the relationship between the controller's cognitive system and the to-be-performed ATC tasks. Thus, the notion of workload proposed by COMETA goes beyond TWL features such as traffic or airspace structure, to include the amount of cognitive effort that is required to successfully perform the task on a context dependent way. COMETA equations attempt to quantify the impact of different events typical of ATC tasks on the complexity of the task. For example, a conflict between two aircrafts generates more complexity for smaller inter-aircraft distances. The air traffic events that are included in COMETA model are typical of enroute sectors (i.e., sectors without take-offs or landings): the intersection between standard routes, the presence of aircrafts in non-standard routes and aircrafts in evolution, as well as the conflicts between pairs of aircrafts. We provide a more detailed description of these equations in the Methods section.

The main objective of MWL models in ATC like COMETA is to predict the best *Demand and Capacity Balancing*, so that they can support ATM planning (Frutos et al., 2019). Particularly, among the ATM functionalities defined in the Pilot Common project adopted by the European Commission (Reg. EU 716/2014), COMETA complexity predictions based on ATCOs mental workload can allow mitigation strategies to be applied at local (ANSPs) and network levels. In general, models for Demand and Capacity Balancing optimization avoid online measures of the ATCOs interventions. This approach to MWL modelling is denominated open-loop because it is based on a-priori information about the traffic, airspace, and operational factors alone to generate predictions of MWL in advance (Loft et al., 2007).

On the other hand, closed loop models analyze MWL as a dynamic interplay between the current TWL state and the actions of the ATCO when dealing with the scenario (Sperandio, 1971; Wee et al., 2018). An example of such models is the one developed by Loft and collaborators (2008). In this model, the ATCO is an adaptive element in the task dynamics, so that MWL is not only a function of the task demands, but it is modulated in feedback loops through the ATCO actions. Research on situational awareness can also be included among these closed loop models (Edwards et al., 2017; Endsley, 1995; Endsley and Smolensky, 1998; Falkland and Wiggins, 2019).

Closed loop models are not currently used to assist Demand and Capacity Balancing prediction because they tend to excessively complicate the process. However, closed loop models can provide more precise and realistic estimations of MWL that could enhance the predictive value of Demand and Capacity Balancing models and potentially improve ATM planning (Kostenko et al., 2016). In this sense, several authors highlighted the need for MWL models in ATC that consider the strategies and action of the ATCOs (Loft et al., 2007).

The original version of COMETA (Frutos et al., 2019) is an open loop model because it does not explicitly include the effects of the actions of the ATCO. In this research, we propose an extension of COMETA model that can account for closed loop situations, and it is more adapted to predict the outcomes of the actual unfolding of ATC tasks. This is achieved by including a dynamic source of MWL generated by the conflicts between aircrafts. The evolution of this conflicts-related MWL is affected by the ATCOs intervention in the aircraft trajectories (note that all the other elements included in COMETA can be obtained a priori from the airspace planning). In addition, this study is the first to externally validate the predictive power of COMETA models using the approach that we explain in the next sections.

### 1.3. MWL model validation

The validation of a MWL model requires as a starting point the rigorous and systematic definition and manipulation of TWL to anchor the predicted variations of MWL produced by the model on a solid ground (Chatterji and Sridhar, 2001; Mitchell, 2000; Reid, 1997; Suárez et al., 2022). Once such relationship has been stablished, we can externally validate a MWL model. As explained above, direct measurement of MWL is not possible. Thus, to validate a model of MWL we require the use of different correlates of workload. Using a diversity of correlates is preferred to enhance the overall quality of the validation process (Young et al., 2015). In this study to validate our model we used the main empirical correlates of MWL identified in the ATC literature that could be measured unobstrusively. These correlates include subjective, performance, and physiological measures, as detailed below.

Subjective questionnaires in the form of rating scales have been widely used because they are easy to fill and they provide direct access to the perceived MWL. In fact, these questionnaires showed a strong relation with TWL and performance (Djokic et al., 2010), and they allow both instantaneous and after-task assessment. As Pagnotta et al. (2021) showed, the most used questionnaires in the field of ATC are the Instantaneous Self-Assessment technique (ISA, Jordan and Brennen, 1992; Tattersall and Foord, 1996), which allows an at-the-moment assessment through a Likert scale, and the NASA-TLX (Hart, 2006; Hart and Staveland, 1988), which assesses the perceived workload through six subscales that are filled after finishing the task. It is important to mention that in the field of ATC, the NASA-TLX is often

used without the subscales weighting procedure, using only the raw evaluation of the six subscales, a procedure denominated NASA-TLX raw (Hart, 2006).

Correlates of MWL based on performance or behavioral measures range from the number of clicks, as an index of the amount of interaction of the ATCO with the system (Shou and Ding, 2013), the number of interventions as changes in direction, speed, or altitude of the aircrafts (Metzger and Parasuraman, 2001), to success indexes like acceptance reaction time or clearances (Tobaruela et al., 2014). However, several studies showed that performance measures are not a direct index of MWL, as ATCOs can have a high level of MWL without an effect on their performance (Sperandio, 1971). The last behavioral measure used to predict MWL is eye-movements (Di Stasi et al., 2010; Marchitto et al., 2016), although this technique was not used in this experiment. The specific set of behavioral and performance measures used in this experiment are described in the Methods section.

The main physiological correlates of MWL are blood pressure, heart rate, breath rate, blink rate, pupil size, and electroencephalograms, as reviewed by Charles and Nixon (2019; see also Pagnotta et al., 2021). The technological evolution of electronic devices has led to wearable sensors that are non-intrusive even with regard to ATC tasks (Nixon and Charles, 2017). In this study we used a waistband to register heart rate (Socha et al., 2020) and electrodermal activity (EDA, Brookings et al., 1996) as most accessible and promising candidate correlates of MWL.

EDA signals are produced by the activity of sweat glands in response to the activation induced by the sympathetic nervous system. This signal can be decomposed in tonic and phasic components that have different time scales and are produced by different processes (Boucsein et al., 2012). The tonic component of EDA signals are slow drifts in the baseline skin conductance of diverse origin, whereas the phasic component is produced by the short time-scale response to external stimuli. In general, the time course of phasic responses involves a rapid rise in skin conductance after the stimulation, followed by an asymptotic relaxation.

Heart rate variability (HRV) is a natural physiological phenomenon by which the time interval between heartbeats varies continuously. A certain level of variability in heart rate is necessary for a healthy cardiac physiology, and reduced levels of HRV has been found to correlate with higher risk of cardiac pathologies (Abildstrom et al., 2003). HRV is measured by the variation in the inter-beat interval (IBI). Using IBI as raw measure, a variety of methods is available to compute different features of HRV (Hughes et al., 2019; Vollmer, 2015). In the context of ATC tasks, research points to a reduction of HRV as a consequence of increased levels of MWL (Aricò et al., 2017; Pagnotta et al., 2021; Radüntz et al., 2021; Socha et al., 2020).

Despite the aforementioned literature, examples are scarce that try to validate predictive models of MWL through a comprehensive assessment of MWL including subjective, performance, and psychophysiological measures. Outstanding exception are Kopardekar and Magyarits (2003), who contrasted several models of dynamic density against the subjective ratings of ATCOs and obtained a $R^2$ of 0.40. In the case of the COMETA model, no external validation has been performed, but ISA estimates have been used to optimize parameters in the model (Frutos et al., 2019). Results showed that such ISA-optimized COMETA model could provide accurate predictions of subjective estimations of MWL (Root Mean Standard Error [RMSE] $>$0.70, average Spearman correlation: 0.85).

Thus, the objective of this work is to perform a comprehensive laboratory study of ATC tasks to externally validate the COMETA model as a predictor of MWL. To that end, we developed a set of ATC scenarios with systematic variations in TWL that constituted the experimental factors of our design: traffic density and airspace complexity. These scenarios simulated enroute ATC situations with varying levels of traffic and airspace complexity. Our hypothesis is that these variations of TWL will induce proportional levels of task difficulty and will produce simultaneous measurable effects on COMETA predictions of MWL and the MWL correlates. Considering the intertwined relationship between traffic density and airspace complexity, we also expect a significant interaction in the MWL in response to their combined effects (the scenarios are described in the Methods section). To assess the representativeness of COMETA as a predictor of MWL, we next perform a multiple correlation analysis between all the dependent variables with significant effects of TWL factors. Finally, we carry out a Principal Component Analysis (PCA) to better comprehend the actual effects of the TWL manipulations used in the experiment.

## 2. Methods

### 2.1. Ethics statement

The local ethics committee approved the experimental protocol (UAM-CEI-110-2163). Participants signed informed consent forms before participating in the experiment.

### 2.2. Participants

A total of 24 university students participated in the study. Their mean age was 20 years (*SD*: 2.1). All of them had normal, or corrected to normal, vision. They had no previous experience with ATC simulation tasks. They were trained to use the ATC simulator during the study, as explained below in the procedure.

## 3. Materials

We used the ATC-Lab advanced simulator (Fothergill et al., 2009). This simulator reproduces a similar interaction to that of real operational environments to the extent that the main forms of interactions and the aircraft performance mimic real ATC environments. The main strength of this simulator is that it allows experimental control to standardize the scenarios and provides timed information of the actions of the participants and the state of the airspace. It includes the radar image of the airspace and allows the main interventions through data-com using keyboard and mouse (see Fig. 1).

The actions allowed in the tested scenarios were aircrafts acceptance, changes in speed and height, and the use of helper information mechanisms such as a ruler and aircraft information labels. The scenarios simulated enroute operations without any operational disturbance in the airspace. The sector was a rectangle of 400 $\times$ 250 nmi (nautical miles). The altitude of the aircraft in the sector ranged between 24,000 and 27,000 ft, although there were no limitations for altitude changes other than the ones imposed by the physics of the aircrafts. The aircraft features were simplified by including only A320 models to facilitate the task because the participants were complete novices in ATC, and, therefore, naïve on the aircrafts features. Aircrafts speed was set at an average of 400 nmi per hour.

Subjective estimations of MWL were measured through two questionnaires. ISA scales (from 1 to 7) were filled every 2 min during the task indicating the number verbally. Raw Nasa-TLX was filled with pen and paper after finishing each scenario. Participants worn an EMPATICA 4E wristband (Empatica Inc., Italy) on the left wrist during the experimental session, providing continuous measures of Blood Volume Pressure (BVP, at 32 Hz sampling rate) and EDA (at 4 Hz sampling rate). EMPATICA software provided estimations of heart rate (at 1 Hz) and IBI using BVP signal.

### 3.1. Design

We implemented a factorial design with two independent factors: airspace and traffic TWL. The traffic TWL factor was the number of aircrafts under control, with two levels: 6 and 12 aircrafts. Across each level, scenarios were designed to maintain the same number of simultaneous aircrafts within the sector. The airspace TWL factor was the structural complexity of the airspace. It was defined on the basis of

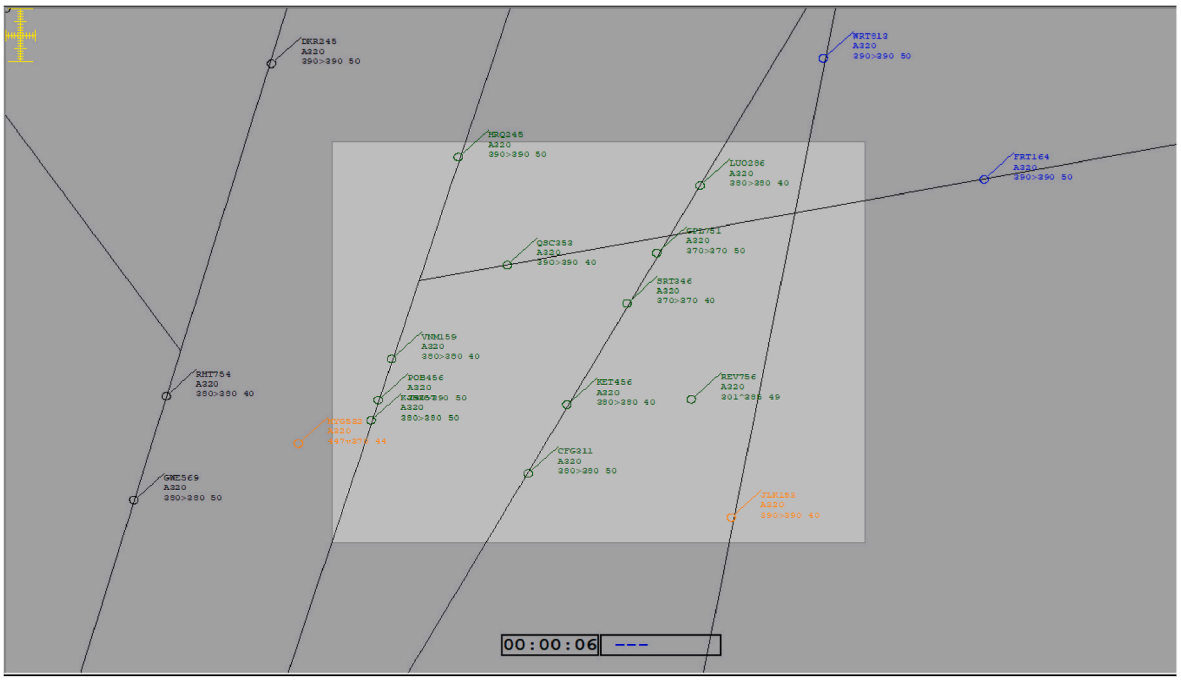


**Fig. 1.** Sample radar image from ATC Lab Advanced.
*Note.* This image illustrates the first seconds of the scenario with high airspace and traffic TWL (see design section for details).

features of the task contemplated by COMETA equations, in particular, the number of standard routes and crossing points, and the number of non-standard routes (to induce more workload, aircrafts in non-standard routes were always in evolution). Airspace TWL had three levels: low, medium, and high.

The differential features of the scenarios are summarized in Tables 1 and 2. The total occupations for all routes presented in Table 2 may be higher in a condition than the total traffic because some routes share segments. In addition, total traffic reflects the number of aircrafts that traverse the sector simultaneously, but the total number of aircrafts in a trial may be higher. A schematic representation of the scenarios is illustrated in Fig. 2. It is important to note that higher levels of airspace TWL imply more routes in which aircrafts are distributed and, thus, a reduced number of interactions between aircrafts for a fixed traffic density. As a consequence, we may expect a reduced or null increase in MWL between medium and high airspace TWL due to the compensatory effect of the increased overall distance between aircrafts, in particular for dense traffic.

We introduced an additional factor in the experimental design to control for the effects of the spatial layout of the airspace. Each scenario was presented in 1 of 4 orientations with rotations of 0°, 90°, 180°, and 270°. As a result, each participant performed the six scenarios produced by the 3 × 2 factorial design in one of the four orientations. The orientations were semi-randomly assigned. Preliminary tests showed that the spatial rotation of the airspace had no effect on the dependent variables. As consequence, to simplify the presentation of the results we have omitted this factor.

**Table 1**
Airspace TWL features of the scenarios.

| Feature | Low | Medium | High |
|---|---|---|---|
| Standard routes (#) | 3 | 4 | 6 |
| Crossing points (#) | 1 | 2 | 3 |
| Non-standard routes (#) | 0 | 1 | 2 |
| Total route intersections (#) | 1 | 4 | 9 |

*Note.* This table summarizes the distribution among the different levels of complexity of task features that contribute more strongly to determine the difficulty of the task. Standard routes indicate routes that are commonly used by aircrafts in a region of the airspace, and are denoted by solid lines in ATClab simulator. Crossing points refer to points of the airspace in which two standard routes cross or merge. Non-standard routes indicate uncommon routes scarcely used and, contrary to standard routes, they are not depicted in ATClab simulator. Total route intersections indicate the total number of intersections between all the routes that cross the sector.

**Table 2**
Occupations of standard route segments in each scenario (aircrafts per simulation time).

| Traffic TWL | Airspace TWL | | |
|---|---|---|---|
| | Low | Medium | High |
| 6 | 8, 7, 5 | 4, 3, 2, 3 | 2, 2, 1, 1, 4, 1 |
| 12 | 9, 8, 10 | 7, 10, 4, 5 | 3, 2, 2, 4, 4, 2 |

*Note.* This table summarizes the number of aircrafts that travel through each of the standard routes defined for each of the three levels of airspace complexity in low and high traffic conditions.

### 3.2. Procedure

The experiment started with a training session that lasted approximately 1 h, in which participants were trained in the use of the simulator. First, they worn the wristband in their left wrist to warm it up, and proceed to read a 10 pages for-the-purpose manual on the functioning of the simulator. Afterwards, they were confronted with six short scenarios in which we tested the correct use of the information and the ATC actions. These included accepting aircrafts in the sector, checking aircrafts, securing separation standards (1000 ft vertical and 5 nmi horizontal), set actions/instructions when needed, and finally hand off the aircrafts to the next sector.

After the training, participants performed six scenarios in a random order. Each scenario took 16 min to complete. Before each scenario, a 60 s baseline measurement with the wristband was performed. Despite the lack of aircraft pilots in ATCLab Advanced simulator (the ATCO can set the instruction directly through data.com), we required participants to verbalize the instructions including the call signs and specific

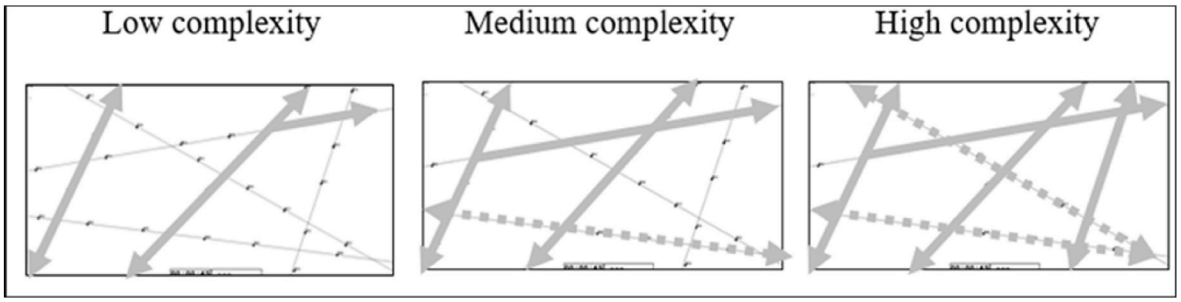


**Fig. 2.** Schematic representation of the airspace TWL levels of the scenarios.
*Note.* Filled lines represent the standard routes whereas dotted lines represent non-standard routes with aircrafts in evolution.

intervention. The researcher, acting as a virtual pilot, responded those verbal instructions with a single word: "received". The whole experimental session lasted for 3 h approximately.

## 4. Theory/calculation

We collected three types of information during a trial: the logs of ATCLab simulator, the subjective estimations of MWL (NASA-TLX and ISA, manually registered in a table), and the physiological correlates of MWL (using EMPATICA wristband). With the information contained in the logs of ATCLab we derived the MWL predictions of COMETA model and the performance variables. We next describe all the dependent variables considered in this experiment and how they were computed.

### 4.1. COMETA model of MWL

COMETA identifies four types of events that can increase the complexity of an enroute sector. The first one is the intersection between two routes. The second event is the existence of aircrafts that travel outside standard routes. The third event is the existence of aircrafts in evolution (i.e., that change their altitude), because the controllers have to use a 2D display to deal with a 3D task. The fourth type of events are the conflicts between aircrafts as an additional source of complexity (i.e. maintain safe distances between aircrafts). These categories were obtained under the guidance of expert ATCOs and their parametrization in the model were optimized in internal tests performed by CRIDA.

COMETA equations attempt to quantify the complexity produced by these multiple events. COMETA considers that these sources contribute additively to the overall complexity faced by the ATCO. Thus, the overall complexity of an airspace situation is simply the summation of all the independent sources of task complexity for all $N$ aircrafts in a sector:

$$COMETA = \sum_{i=1}^{N} 1 + w_{\mathrm{fl}} \bullet C^{i}_{\mathrm{fl}} + w_{\mathrm{cf}} \bullet C^{i}_{\mathrm{cf}} + w_{\mathrm{ev}} \bullet C^{i}_{\mathrm{ev}} + w_{\mathrm{ns}} \bullet C^{i}_{\mathrm{ns}} - C^{i}_{\mathrm{rd}},$$

where $w_{\mathrm{fl}} = 1$, $w_{\mathrm{cf}} = 4$, $w_{\mathrm{ev}} = 2$, $w_{\mathrm{ns}} = 2$, are the relative weights of the four sources of complexity considered by COMETA: standard flow interactions ($C_{\mathrm{fl}}$), aircraft conflicts ($C_{\mathrm{cf}}$), aircrafts in evolution ($C_{\mathrm{ev}}$), and aircrafts in non-standard routes ($C_{\mathrm{ns}}$). The values of these weights were selected to equalize the scale of the four factors, without performing an a priori optimization. The last factor ($C_{\mathrm{rd}}$) is a reduction of the minimal complexity of 0.5 applied to aircrafts in which all the other factors are zero.

These components of complexity are first computed for each of the $N$ aircrafts in the scenario and then added together. We can also compute the overall complexity of each source independently. COMETA and its components were computed every second (the update rate set on ATCLab), although for the statistical analyses we used the average of these time series. We next describe the computation of the components of COMETA. For each of the sources of complexity, the model first defines an intermediate variable called severity, from which the final value of task complexity is computed applying different thresholds on the severities, as indicated for each source below.

*Standard Flow interaction-related complexity (Cfl):* the severity associated to the interaction between two standard flows is:

$$S_{\mathrm{fl}} = \mathrm{T} \bullet \mathrm{O}_1 \bullet \mathrm{O}_2$$

where T is an interaction complexity factor and $O_1$ and $O_2$ are the respective occupations of the interacting routes. The severity of a flow interaction is proportional to the number of aircrafts that travel through this route per unit of time. In addition, the factor T represents the increase in severity due to the respective altitude changes of the two interacting routes. Thus, $T = 1$ if none of the routes in interaction is evolving; $T = 2$ if only one route is evolving; $T = 3$ if both routes are evolving in the same sense; and $T = 4$ if the two routes are evolving in opposite senses. The complexity results from applying the thresholds $i_1 = 32$ and $i_2 = 25$ to the severity obtained for each flow interaction: if $S_{fl} > i_1$ then $C_{fl} = 0.2$; if $S_{fl} < i_1$ and $S_{fl} > i_2$ then $C_{fl} = 0.1$; and if $S_{fl} < i_2$, $C_{fl} = 0.05$.

*Conflicts-related complexity ($C_{cf}$)*: A conflict is defined as a potential crossing between the trajectories of two aircrafts. A conflict is considered active when it occurs within the following spatiotemporal boundaries: the intersection between the trajectories has to occur in less than 10 min in the future and inside the sector, the vertical distance between the trajectories in the intersection must be lower than 800 ft, and both aircrafts have to be within 10 nmi of the intersection. COMETA model considers four factors that contribute to conflict severity:

- *A1* is the horizontal separation in meters between the aircrafts involved in the conflict at the time of calculation.
- *A2* is the proximity of the crossing point to the frontier of the sector: if distance <10 nmi, then $A2 = 1.02$, else $A2 = 0.90$.
- *A3* is the angle of convergence between the trajectories in the conflict point: if angle <90°, then $A3 = 1.10$, else $A3 = 0.80$.
- *A4* is the proximity of the conflict to sector's critical points. If distance <1500 m, then $A4 = 0.90$, else $A4 = 1.05$.

We obtain the overall severity of a conflict as the product of all four factors:

$$S_{cf} = A1 \bullet A2 \bullet A3 \bullet \mathrm{A4}$$

Note that conflict severity is inversely defined, that is, higher values of severity indicate less complexity. $C_{cf}$ is then computed using the thresholds $c_1 = 9260$ and $c_2 = 4630$. If $S_{cf} > c_1$, then $C_{cf} = 0.1$; if $S_{cf} < c_1$ and $S_{cf} > c_2$, then $C_{cf} = 0.2$; if $S_{cf} < c_2$, then $C_{cf} = 0.3$.

*Aircrafts in evolution-related complexity ($C_{ev}$)*: When an aircraft is changing the altitude, $C_{ev} = 0.10$.

*Non-standard-related complexity ($C_{ns}$)*: In aircrafts that traverse a non-standard route, $C_{ns} = 0.15$.

These complexity variables are computed every second for every aircraft in the sector. The overall measures of complexity are obtained by adding the individual values. In the results sections we consider the

overall COMETA index as well as each of its five components.

### 4.2. Subjective variables

ISA and raw NASA-TLX subjective estimations were registered as explained in the procedure. An averaged ISA value was computed for each trial, and the overall NASA index was computed as the sum of all its components.

### 4.3. Physiological variables

An EMPATICA wristband was used to measure heart rate and EDA. The physiological time series provided by these devices had too much lost data due to long periods in which the signals had artifacts. We detail below the specific processes used to improve the physiological signals, as well as the dependent variables included in the analysis.

We used the cvxEDA model to separate both components from our signals (Greco et al., 2015). Before submitting each time series to the cvxEDA model, we first smoothed the signal using a 4th-order low pass Butterworth filter with a cutoff frequency of 0.9 Hz (Subramanian et al., 2019). After separating tonic and phasic components using cvxEDA function, we normalized them dividing by the average value of the respective component, obtained during the 60 s baseline measurement before each trial. The dependent variables submitted to factorial analysis were the normalized per-trial averages of $EDA_{phasic}$ and $EDA_{tonic}$.

The registration issues of EMPATICA wristband were particularly severe for the heart rate. As a consequence, we used the raw signal provided by the device, BVP, to increase the amount of valid data collected. The first problem of the BVP signals was that the amplitude of the cycles varied heavily during a trial. To solve this issue, we performed a dynamic range compression (using MATLAB's function *compressor*) to homogenize the signal gain for each trial. The second issue of the BVP signals was the abundance of high frequency noise, as well as very slow oscillations. As we were only concerned with variations in BVP that indicated heart beats, we used a 4th-order band pass Butterworth filter to remove frequencies below 0.7 Hz and above 2 Hz. The IBIs obtained from a peak detection algorithm (function *peakdet* from MATLAB) applied to this smoothed signal were further scrutinized to remove data points that were not physiologically plausible. The exclusion criteria for the detected beats were: more than 220 beats per minute or less than 20; variations of beat time higher than 20% from the previous beat, or longer than 300 ms. Finally, we used the interquartile range method to exclude outliers (IQR; Vollmer, 2015). From the set of IBI remaining after the filtering we computed the average heart rate and the SDNN (Standard Deviation of normal to normal R-R intervals; this is the most basic measure of HRV, other more complex measures of heart rate variability were considered, but we omit them because the results were similar to these variables).

### 4.4. Performance and behavioral variables

We measured the following dependent variables to estimate either the outcome of the control actions or the control actions themselves. The variable *Conflicts Reduction* represents the reduction in the average number of conflicts during a trial with respect to the conflicts that occur in a passive simulation of the same scenario (without any intervention from the participant). We used the COMETA definition of conflict detailed above. Best ATCO performance is assumed to produce a significant reduction of the conflicts with respect to not doing anything in the passive simulation. The *Centroid Distance* captures the average spatial dispersion of the aircrafts in the sector. It is a complex variable affected by TWL factors such as the routes layout, the traffic density, as well as by the actions of the controller. To compute it, we first obtained the centroid of the positions of the aircrafts at each time step. Then, we computed the average distance to this point. In general, for a given scenario, performance is better when the distance between aircrafts is maximized. The *Speed Success* and *Altitude Success* indicate how closely the participants followed the flight plan prescribed for each aircraft. They are defined as the number of aircrafts that left the sector having the speed or altitude prescribed in the flight plan. Finally, we measured four variables to estimate how the participants were executing the task. First, the *Accept Reaction Time* indicates the average time required by a participant in a trial to accept aircrafts after they announce themselves. Longer reaction times tend to indicate a higher workload, although it might depend on the ATCO strategies. The *Total Clicks* variable is a measure of the total amount of interactions with the simulator performed by participants during the trial. We assume that higher TWL is related with larger amount of mouse clicks, but it may be too sensible to the idiosyncrasy of each participant. The variables *Altitude Interventions* and *Speed Interventions* measure the total amount of altitude and speed changes in the aircrafts respectively performed by the participant in a trial. These variables are affected by the amount of MWL, but also by the strategy used by the participants. The preferred strategy would be one that minimizes the number of interventions, although in a simulation study with novices it is not uncommon to observe more interactive strategies that still manage to keep conflict severity low.

### 4.5. Statistical analysis

All dependent variables were submitted to repeated measures analysis of variance (ANOVA) using as main factors airspace TWL (low, medium, high) and traffic TWL (6, 12). A preliminary ANOVA of the sequential ISA measurements of each trial was performed using an additional Time factor (with levels 2, 4, 6, 8, 10, 12, 14 and 16). Due to the lack of effect of factor Time, we averaged ISA measures for each trial and report only the results from this averaged variable. Huynh-Feld corrections were used when sphericity assumptions were not met. Post-hoc analyses were performed with Tuckey tests. Multiple correlation and principal component analysis were carried out using *xcorr* and *pca* functions from MATLAB, respectively.

### 4.6. Software tools

All the data processing and modelling were performed using self-developed software. We have developed *pyatc* python package to process the output of ATCLab log files and compute COMETA and performance variables. The version of pyatc used in this research can be obtained in https://github.com/jibannez/JATM_2022_pyatc. The analysis of the time series was performed with self-developed MATLAB code. The statistical analyses were performed with *jamovi 2.8* (The jamovi project, 2022).

## 5. Results

### 5.1. ANOVA of COMETA model

Fig. 3 summarizes the group means and standard errors for the overall COMETA index and its five components: COMETA Flow, COMETA Conflict, COMETA Evolution, COMETA Non Standard, and COMETA reduction. ANOVA analysis of COMETA model revealed a strong significant interaction between airspace and traffic TWL factors, $F(2, 46) = 52.50$, $p < .001$, $\eta_p^2 = .695$. In general, higher values of both airspace and traffic TWL produced higher values of COMETA. Posthoc analyses showed that only the comparison between medium and high airspace TWL with 12 aircrafts was not significant. All the remaining paired comparisons were significant. As already advanced in the methods sections, this was expected because high airspace TWL implies that more routes are available to host the same number of aircrafts, producing a dispersion of traffic that reduces the overall complexity of the trial.

The ANOVA of COMETA Flow indicated a strong significant interaction between airspace and traffic TWL, $F(2, 46) = 1368$, $p < .001$, $\eta_p^2 =$

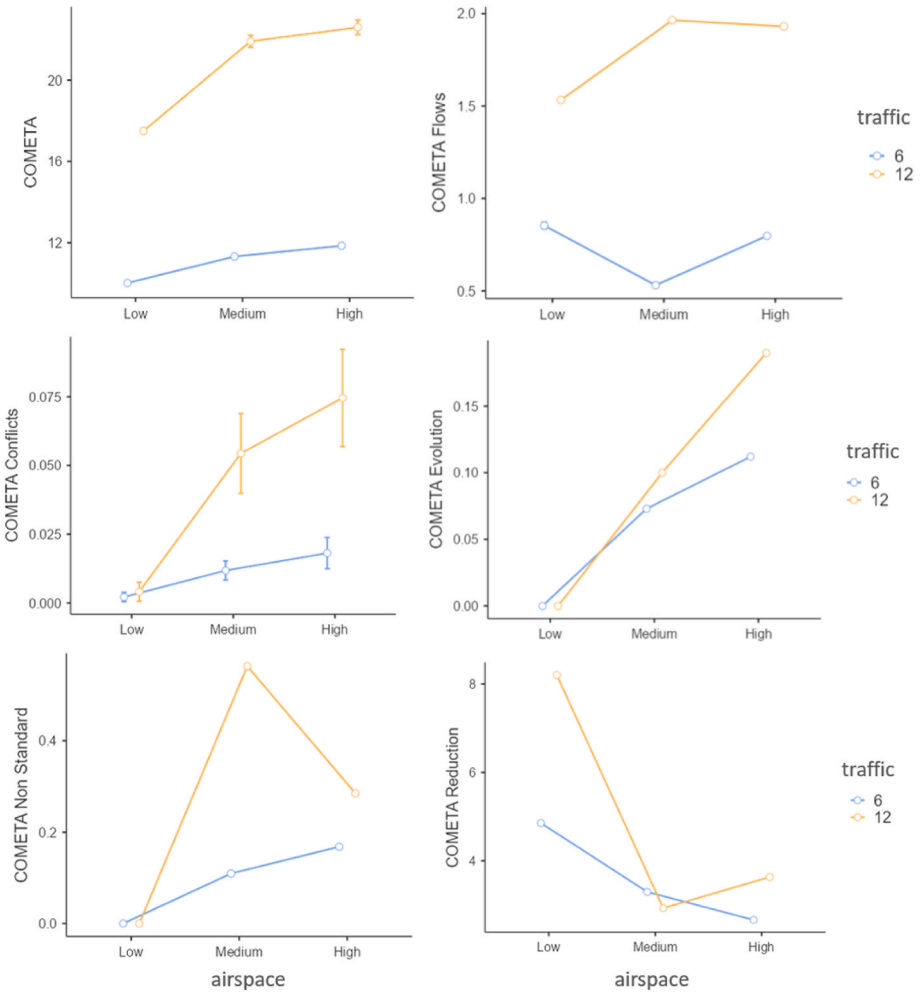


**Fig. 3.** Means and standard errors of variables from COMETA model as a function of TWL factors.
*Note.* All the dependent variables represented are dimensionless indexes described in detail in the Methods section.

.983. Posthoc analysis revealed that all paired comparisons were significantly different with the single exception of low and high airspace TWL with low traffic. To understand this effect, we must consider that COMETA Flow is computed for all aircrafts that traverse a standard route, and that the occupations of the routes have a multiplicative effect on the final value of complexity. If we recall the occupation coefficients presented in Table 2, the reason for this effect is much clearer. Multiplying the occupations of each condition one can observe that the condition with medium airspace TWL and 6 aircrafts has the lowest product. A similar mechanism explains the reduction of MWL between scenarios with medium and high airspace TWL when the traffic is dense.

The interaction between airspace and traffic TWL was significant for COMETA Conflict, $F(1.70, 39.05) = 5.23$, $p = .013$, $\eta_p^2 = .185$. Posthoc analysis revealed that all level combinations with low traffic are not significantly different. In addition, COMETA Conflict was also not significantly different when comparing traffic conditions at low airspace TWL. These results indicate that the traffic density on conditions with only 6 aircrafts was too low to produce conflicts, regardless of the airspace structure. When the traffic density was higher, airspace structure was again relevant: More complex structures produced higher conflict-related MWL.

COMETA Evolution closely replicated the number of aircrafts in evolution present in each condition. Thus, the interaction between airspace and traffic TWL was significant, $F(2, 46) = 1.86 \bullet 10^6$, $p < .001$, $\eta_p^2 = 1$. All paired comparisons were significant in the post hoc tests, with the exception of the two low airspace conditions without aircrafts in evolution. A similar relationship can be observed in the results of COMETA Non Standard, which closely follows the number of non-standard aircrafts present in each condition. Thus, the interaction between airspace and traffic TWL was again significant, $F(1.02, 23.44) = 5.44 \bullet 10^6$, $p < .001$, $\eta_p^2 = 1$. All paired comparisons in the posthoc analysis were significantly different.

COMETA Reduction is meant to compensate for those aircrafts that are present in the sky but have no complexity associated with them from the other components. Thus, its value is higher for these conditions with less density or with more heterogeneously distributed events. ANOVA showed a very strong effect of the interaction between experimental factors, $F(1.01, 23.31) = 5198$, $p < .001$, $\eta_p^2 = .996$. All paired comparisons in the posthoc analysis were significantly different.

### 5.2. ANOVA of subjective scales

Group averages and standard errors of subjective measures of MWL are illustrated in Fig. 4. ANOVA of ISA scores revealed a significant main effect of the dynamic TWL factor, $F(1, 23) = 74.50$, $p < .001$, $\eta_p^2 = .764$. The main effect of the factor airspace TWL was also significant, $F(2, 46) = 10.64$, $p < .001$, $\eta_p^2 = .316$. The interaction between the main factors was not significant, $F(2, 46) = 0.56$, $p = .570$, $\eta_p^2 = .024$. Posthoc analysis evidenced that all paired comparisons between low traffic and high traffic were significant for the same airspace structure. In addition, the comparison between low and high static TWL was also significantly different for the same traffic level. In sum, the ISA scores were linearly

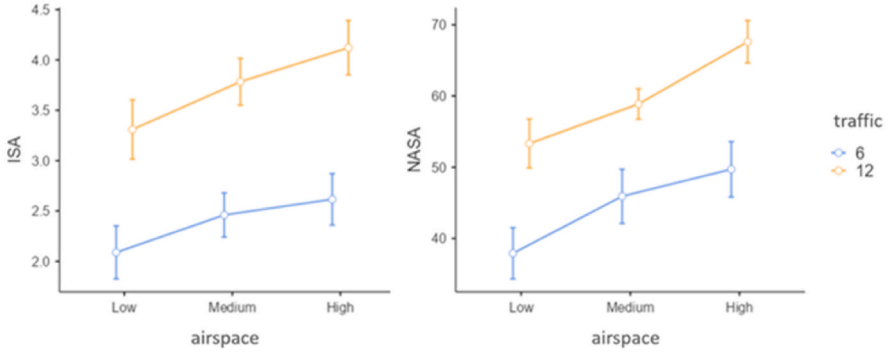


**Fig. 4.** Means and standard errors of subjective estimations of MWL as a function of TWL factors.
*Note*. These variables are dimensionless indexes.

increasing in response to increases in airspace and traffic TWL independently.

The results from NASA scores were compatible with the above discussed results from ISA scores, with a significant main effect of the two factors. For the airspace TWL, $F(1, 23) = 38.90$, $p < .001$, $\eta_p^2 = .628$. For the traffic TWL, $F(2, 46) = 23.90$, $p < .001$, $\eta_p^2 = .510$. The interaction between the two factor was not significant, $F(2, 46) = 0.85$, $p = .430$, $\eta_p^2 = .036$). Posthoc analysis revealed the same pattern as in ISA scores: the perceived MWL was linearly and independently increasing in response to increases in airspace and traffic TWL.

### 5.3. ANOVA of performance measures

Group averages and standard errors of performance measures are illustrated in Fig. 5. The ANOVA of Conflicts Reduction had a significant interaction between airspace and traffic TWL, $F(2, 46) = 307$, $p < .001$, $\eta_p^2 = .930$. Posthoc analysis evidenced that, with the exception of the comparison between the two traffic levels at the simplest airspace structure, all the remaining comparisons were significantly different. This indicates that for simple airspaces, there was no possible reduction in the number of conflicts. For more complex scenarios, the participants managed to reduce the number and duration of conflicts. This reduction was much more intense for higher traffic densities.

The ANOVA of Centroid Distance showed a significant interaction between airspace and traffic TWL, $F(1, 23.05) = 29.20$, $p < .001$, $\eta_p^2 = .560$. All paired posthoc comparisons were significant. This variable is determined by the spatial layout of the airspace and the relative occupation of the different routes. Lower values indicate that the aircrafts are spatially more concentrated. This explains why the medium level of airspace TWL had the lowest centroid distance, as it has a complex airspace structure with a relatively low number of standard flows.

The ANOVA of Altitude Exit Success evidenced a significant interaction between traffic and airspace TWL, $F(1.34, 30.89) = 15.90$, $p < .001$, $\eta_p^2 = .409$. Posthoc analysis revealed that for low and high airspace TWL, the differences between traffic conditions were significant. In particular, we observed less success ratio for dense traffic in the low airspace TWL condition, and the opposite tendency for high airspace TWL. In medium airspace TWL the success ratio was similar for the two traffic densities. The ANOVA of Speed Exit Success closely replicated the results of the variable Altitude Exit Success, with a significant interaction between traffic and airspace TWL, $F(2, 46) = 69.38$, $p < .001$, $\eta_p^2 = .268$. Posthoc analysis again replicated the pattern observed in Altitude Exit Success. These results suggest that participants failed to comply with the flight plans, and proportionally more intensely for scenarios with higher TWL.

ANOVA of Altitude Interventions showed a significant interaction between airspace and traffic TWL, $F(2, 46) = 8.44$, $p < .001$, $\eta_p^2 = .268$. Posthoc analysis evidenced that significant paired differences were only found for conditions with high airspace TWL. Participants used more altitude interventions with increasing traffic density and airspace complexity. On the contrary, the ANOVA of Speed Interventions could not find any significant effects. Fig. 5 clearly ilustrates the mostly flat and fluctuating response of this variable with respect to the experimental factors.

None of the main effects or interaction were significant in the ANOVA of Accept Reaction Time. Despite the tendency that can be observed in Fig. 5 to have longer reaction times for more dense traffic or more complex airspace structures, the large variability in all conditions diluted these central tendencies (although the main effect of traffic TWL was very close to the critical alpha value).

The ANOVA of Total Clicks evidenced significant main effects of airspace TWL, $F(1, 23) = 45.31$, $p < .001$, $\eta_p^2 = .663$. The main effect of traffic TWL was also significant, $F(2, 46) = 27.73$, $p < .001$, $\eta_p^2 = .547$. The interaction was, on the other hand, not significant. Posthoc tests evidenced that all paired comparison between traffic conditions with the same level of airspace TWL were significantly different. The comparisons with the same level of traffic TWL were significantly different between low airspace TWL and both medium and high levels, but not between the two higher levels of airspace TWL. This means that the amount of interactions required saturated after the medium airspace TWL, and was proportional to the traffic density.

### 5.4. ANOVA of psychophysiological measures

As can be seen in Fig. 6, all pychophysiological measures had a similar flat response to the design factors, with fluctuations larger than the differences between conditions. The only marginally significant effect evidenced by ANOVA was the interaction between main factors for $EDA_{phasic}$, $F(2, 46) = 2.47$, $p = .096$, $\eta_p^2 = .097$. This result is produced by the markedly higher value of phasic activation in the condition with middle airspace TWL and high traffic density. This condition was expected to be the most challenging to perform according to COMETA prediction.

### 5.5. Correlational analysis

We performed multiple correlation analysis between the dependent variables that had significant effects in the repeated measures ANOVA. The results are summarized in Table 3. COMETA index had high correlation with all the variables included in the analysis (with the exception of COMETA reduction), which indicates that it is a good index of MWL. This correlation was particularly strong with COMETA Flow and COMETA Non Standard, which indicates that the static properties of the airspace in our scenarios were the most relevant to determine COMETA index. It is important to note that this is not a general property of COMETA, but a contingent feature of the TWL manipulations used in our design.

The correlation of COMETA with the performance variables Conflicts

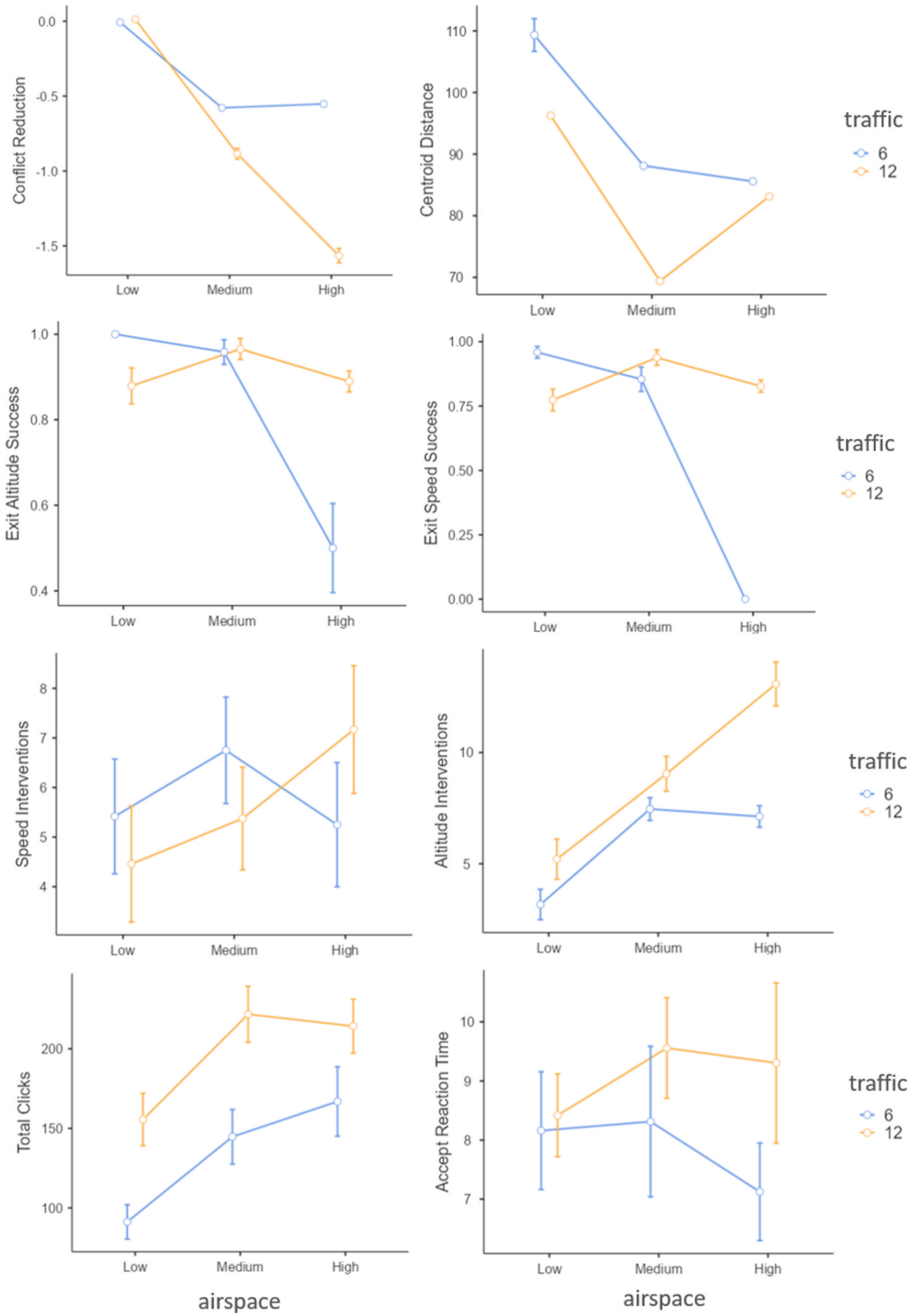


**Fig. 5.** Means and standard errors of performance and behavioral correlates of MWL as a function of TWL factors.
*Note.* All the dependent variables represented are dimensionless indexes described in detail in the Methods section.

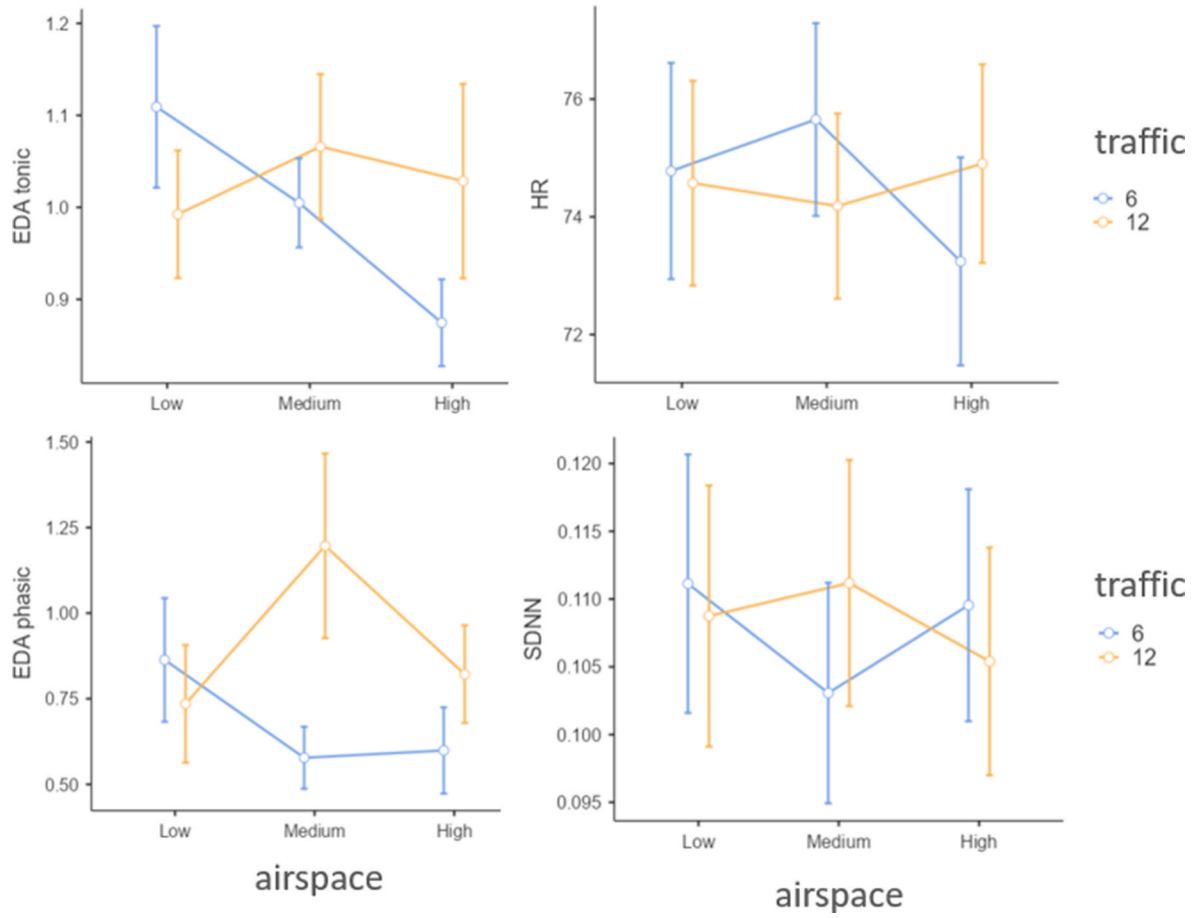


**Fig. 6.** Means and standard errors of psychophysiological correlates of MWL as a function of TWL factors.
*Note*. We have included results from EDA, HR and HRV variables although none of them had any significant effect. The lack of significant differences is due to the much larger variability in each condition than between conditions. We interpret this effect as a measurement issue of the EMPATICA wristbands.

**Table 3**
Correlations for study variables.

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. COMETA | | | | | | | | | | | | | | |
| 2. C. flow | .95** | | | | | | | | | | | | | |
| 3. C. conflict | .57** | .38** | | | | | | | | | | | | |
| 4. C. evolution | .54** | .39** | .44** | | | | | | | | | | | |
| 5. C. non standard | .67** | .63** | .40** | .62** | | | | | | | | | | |
| 6. C. reduction | .02 | .14 | −.23 | −.67** | .59** | | | | | | | | | |
| 7. Conflict reduction | .68** | .56** | .47** | .93** | .69** | −.55** | | | | | | | | |
| 8. Cent distance | −.61** | −.50** | −.34** | −.63** | −.84** | .52** | −.63** | | | | | | | |
| 9. Altitude Success | .13 | .13 | .09 | −.17 | .04 | .15 | .01 | .08 | | | | | | |
| 10. Speed Success | .27** | .30** | .09 | −.22* | .10 | .25* | .05 | .10 | .62** | | | | | |
| 11. Altitude interv. | .49** | .39** | .35** | .62** | .43** | −.31** | .65** | −.47** | −.03 | −.09 | | | | |
| 12. Total Clicks | .40** | .36** | .17* | .36** | .40** | −.16 | .36** | −.41** | −.12 | .01 | .15 | | | |
| 13. ISA | .50** | .48** | .23* | .31** | .34** | .00 | .38** | −.36** | .11 | .12 | .41** | .03 | | |
| 14. NASA | .46** | .40** | .31** | .37** | .31** | −.06 | .40** | −.35** | .06 | .00 | .34** | .18* | .75** | |

*Note.* Abbreviations used in the table: C. (COMETA), interv. (interventions). *$p < .05$. **$p < .01$.

Reduction and Centroid Distance was also strong. These two variables are measures of the dynamic outcomes of close-loop control, which indicates that adding conflict related complexity to the model successfully captures part of these online dynamics. Finally, COMETA had significant correlations with the two subjective measures of MWL, ISA and NASA. In fact, the correlation with COMETA was the largest for these variables with the exception of their mutual correlation. This indicates that the estimations of MWL from COMETA were compatible with the subjective estimations provided by the participants (despite the above explained discrepancy).

In the remaining correlations there are some interesting results that clarify aspects of the complex interactions that emerge in the task. For example, there is a large correlation between the triplet COMETA Flow, COMETA Non Standard, and COMETA Evolution. This is an expected effect considering that these three variables are linked by the definition of the airspace TWL experimental factor (see Table 2). In addition, the variable COMETA Reduction had the strongest correlation with the variables COMETA Evolution and COMETA Non Standard, because in our scenarios these two variables determine COMETA Reduction: When there are non-standard aircrafts (which are always in evolution),

COMETA Reduction is always zero.

Conflict Reduction had strong correlations with all performance variables (except Exit Altitude Success, which had low correlations with all variables except Exit Altitude Speed), indicating that performance can be conceived as a whole in which a participant is either having a good overall performance or not. In addition, Conflict Reduction and Centroid Distance had strong correlations with COMETA Evolution and COMETA Non Standard, indicating that the actions of participants were particularly effective for scenarios with non-standard aircrafts.

### 5.6. PCA analysis

As a final step in our analysis, we wanted to highlight to what extend the responses of the dependent variables expressed the factorial manipulations in TWL of our experimental design. Considering that all the dependent variables represent different approaches to the measurement of MWL, this PCA attemps to validate the manipulations on TWL in the scenarios used. To that end we performed a PCA analysis in which the variables were the 6 combinations of airspace and traffic TWL experimental levels (A: low-6; B: medium-6; C: high-6; D: low-12; E: medium-12; F: high-12), and the cases were the dependent variables of the design that had significant effects in the ANOVA (the same subset used in the previous section). We averaged the values of each variable over the 24 participants to obtain a matrix with 6 rows and 14 columns. Scores from the non-averaged dataset were similar but are more complex to interpret. Thus, we present the results on the participant-averaged dataset. Before submitting the matrix to the PCA, we normalized column-wise, substracting the mean and dividing by the standard deviation of each column.

Fig. 7a includes the scores obtained by the six variables in the first two principal components that explained 85% of the variance. There we can see that the first principal component (the horizontal dimension of the plot), which explained 60% of the variance, separated scenarios with respect to their traffic density. The only exception was scenario D (12 aircrafts with low airspace TWL) that had a similar score as scenarios B and C. On the other hand, the second principal component (the vertical dimension of the plot), which explained 20% of the variane, separated the different scenarios according to their structural complexity. In this case, the separation correctly organized all the different scenarios, but the projections differed between traffic conditions: low density scenarios spaned positive and negative values, whereas high traffic scenarios were confined to the first quadrant, spanning up to three times less variability in this dimension. Taken together, these results suggest that scenario D was closer to a low TWL condition, and that the TWL of scenarios E and F was too close to be funcionally distinguishable.

Fig. 7b presents the relative contributions of the dependent variables in the determination of the principal components. To provide an interpretation of their role we focus on those variables with a weight larger than .25 in absolute value. Thus, the first component was dominated by Centroid Distance in the negative side (so larger values of the variable produce more negative projections), whereas the positive side contained a large number of variables with very similar weights (around 0.3). This indicates that the variability of the original dataset was equally distributed among all these variables. With respect to the second principal component, the situacion is much clearer. This component was dominated by a pool of closely related variables: Exit Altitude and Speed Success, COMETA Evolution, and COMETA Reduction. As already explained, these variables were tightly coupled due to the specific design of this experiment. In other words, our design of the structural complexity of the routes could be sufficiently described by referring only to the number of aircrafts in evolution.

## 6. Discussion

The main aim of this study was to propose and validate a model to predict ATCO MWL. To that end, we adapted COMETA model (an open loop model that only requires data about traffic plans and do not consider the actual interactions of the operator), to a closed loop situation such as an operator performing an ATC task. To validate the model we conducted a laboratory experiment with an ATC simulator in which we systematically varied features of the scenarios that are known to affect TWL. In addition, we measured correlates of MWL to validate the predictions of our model. The results demonstrate that our model can predict the increase of MWL produced by increases in airspace complexity and traffic density, as well as their interactions. This general pattern of results is reproduced by the subjective responses and the performance and behavioral correlates. Let us explain in more detail the results.

The experimental results demonstrate that COMETA predicts changes in MWL generated by changes in traffic and airspace structure. COMETA components are also sensitive to TWL manipulations that affect the source of MWL represented by each component. COMETA predicts higher increments of MWL due to traffic factors than air space structure. In fact, traffic load measure in the form of dynamic density metrics has been shown to be the main determinant of TWL (Kopardekar and Magyarits, 2003), and is a basic tool for controlling the sector load in air traffic. This is reasonable to expect because a complex structure with low traffic brings about a very simple ATC task. This is implemented in COMETA equations because the traffic density affects all sources of MWL. Nevertheless, the COMETA model has been also sensitive to airspace structure factors, bearing out the structural approach to cognitive complexity proposed by Histon and Hansman (2008).

COMETA predictions of MWL allow us to better understand and optimize the experimental manipulations on TWL. Our results indicate

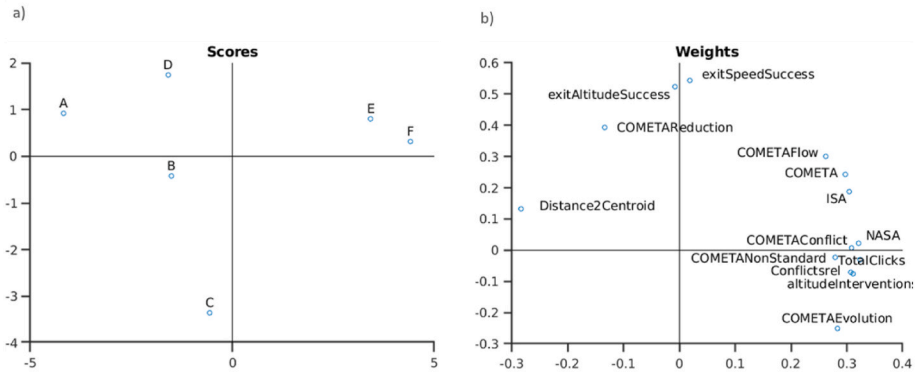


**Fig. 7.** Results of the PCA analysis over the TWL factors and dependent variables.
*Note.* Panel A shows the scores obtained by the six scenarios in the first two principal components. Panel B shows the weights on the two first components for those variables with absolute values weights larger than 0.25.

that COMETA successfully predicts that for the scenarios with the simplest structure we cannot obtain different TWLs for traffic densities, because the task is too easy. In the same vein, COMETA predicts that a proliferation of routes increases the airspace TWL, but this increase may not have an impact on MWL because it also decreases the number and severity of the interactions between standard routes and aircrafts, as they spread out in the different routes increasing the centroid distance of the scenario. Previous research has shown that minimum separation between aircrafts strongly affects performance and perceived complexity (Boag et al., 2006). This is evident in the experimental scenario with high traffic and medium airspace TWL, which produced the lowest centroid distance and resulted with a MWL prediction close to the high-high TWL scenario.

Subjective estimations of MWL increase linearly with increases of the two TWL factors, without interaction between them. Thus, participants' reports did not reflect the interactions predicted by COMETA and explained in the previous paragraph. Our hypothesis is that these discrepancies can be explained taking into account performance and behavior, as these are the most accessible handles that the participants have to produce subjective estimations. However, the subjective reports also show the larger differences in MWL produced by traffic compared to those produced by airspace structure.

The general trend of performance and behavioral measures replicates the tendency predicted by COMETA model with a saturation for scenarios with medium and high static TWL with dense traffic, as well as well the lack of discrimination for too easy tasks. In particular, this is the case for Centroid Distance, Exit Speed Success, Exit Altitude Success, Total Clicks and Accept RT. However, there are two notable exceptions to this trend. Conflicts Reduction and Altitude interventions present a very strong effect of the airspace structure, even more intense for scenarios with high airspace TWL. This tendency contradicts COMETA predictions and parallels the subjective estimations of participants.

These results indicate that the weight of conflicts on COMETA predictions for our scenarios is relatively low, which could explain the difference between COMETA predictions and subjective judgements if participants preferentially use conflict management actions to estimate MWL or other context dependent features of the scenarios (Colle and Reid, 1998). Note that this effect is not a necessary property of COMETA model, but a contingent feature of the set of scenarios used in this experiment.

Correlational analysis of the relationships between the dependent variables confirm several of the results obtained in the factorial analysis. First, COMETA is the only variable that has high correlation with all the other indexes of MWL. This demonstrates that COMETA is a good index of the multifactorial nature of MWL in ATC tasks. In particular, COMETA has the strongest correlation with subjective estimates, which have been shown to be the most faithful access to MWL (Djokic et al., 2010). Second, the airspace sources of TWL have the highest correlation with COMETA predictions in this experiment. As explained above, this is contingent to the set of scenarios used, and could be mitigated if a different selection of features is used. Third, the high correlation between COMETA and performance variables is probably mediated by the strong correlation of performance variables with airspace sources of TWL. Relatedly, the three components of COMETA that are directly related with airspace TWL (Flows, Evolution, and Non Standard) are strongly correlated. This is an expected result considering how we defined this factor (see Table 2). And finally, we have observed that performance and behavioral variables tend to correlate internally, which indicates that we can treat performance as a whole: a participant is either having good or bad performance, regardless of the dimension tested.

To conclude we performed a PCA to systematize the effects of the TWL manipulations used to obtain insights for future developments. This analysis has shown that the two TWL factors defined in this experiment were strongly coupled, but were still remarkably good to represent a range of different air traffic situations that resemble natural environments. The first principal component separated low traffic from high traffic scenarios (with the exception of D scenario, which was too simple). This component was determined by the Centroid Distance on one side of the dimension, and by all the remaining variables in the design on the other. All variables contributed equally to this component, which reflects the global effect that traffic density has on ATC tasks. This result is in support to those approaches that rely on density measures as an index of complexity (Kopardekar and Magyarits, 2003). The second principal component separated and organized the scenarios according to their structural complexity. This component was determined by variables related to the number of aircrafts in evolution, which reflects the contingent features of our set of scenarios.

Finally, the signals from the psychophysiological correlates measured were not conclusive. We interpret this general lack of results as caused by the convergence of three factors. First, the tasks were not particularly stresfull for the participants, with only the most engaged ones showing enough sustained concentration to expect an arousal. Second, the slow pace of the potential stressors aggravated the lack of tension that could be expressed in psychophysiological measures. Third, the quality of the measures was too low, probably due to a fundamental limitation of the hardware reliability or its sensitivity to variations on MWL levels (Mach et al., 2022).

## 7. Conclusions

This study demonstrates that the proposed model can successfully predict MWL in a range of environments. There is room for improvements but this model has promising future applications in ATC planning. Particularly, COMETA can participate into the Network Collaborative Management ATM functionalities defined in the Pilot Common Project derived from the SESAR R&I solutions (EU, 2014). Thus, COMETA may assist on the design of mitigation strategies to be applied at local (ANSPs) and network levels including the ATCOs MWL in complexity prediction.

The introduction of conflict related complexity in COMETA model represents a first step in the integration of ATCOs interactions in air traffic planning. It still remains a very complex task to perform, but we have provided an illustration of the path required. For example, we have defined a reference state that consisted in running the ATC task without any intervention. This reference state enables us to evaluate performance and MWL with respect to the improvement (or worsening) produced by the ATCOs actions. The integration of close loop models in ATM would certainly require the development of sophisticated interactive patterns that capture the strategies and mean-field behavior of ATCOs.

There are three main shortcomings in our study. The first one is the use of novice participants that were specifically trained for this study, and the use of adapted tasks in the ATC simulator. Despite our encouraging results, field studies with real tasks and ATCOs are necessary to fine-tune the parameters of the model. The second main shortcoming has been the poor results from the psychophysiological correlates of MWL, as the original plan if the measured signal had been good markers of MWL was to first optimize COMETA parameters with them, to later perform the external validation with subjective measures. Thus, a future development of this research must be to obtain a good physiological marker of MWL in ATC tasks. Finally, our model does not consider the possibility of changes in the strategy to solve the task nor the reciprocal relationship between situational awareness and MWL. These are among the most relevant factors to understand ATCO performance in the real world, so future developments in MWL modelling must include an account of them.

We have provided a systematic method to manipulate TWL inspired in the analysis of the task done by COMETA. Although this approach to study ATCO performance has been successful to provide an objective and measurable anchor to MWL models, TWL has demonstrated to be highly complex in itself. Thus, the complexity of TWL should be

independently and quantitatively explored as a precondition to better models of MWL (EUROCONTROL, 2004; Histon and Hansman, 2008).

Finally, the pervasive and well known effect of traffic density on all aspects of both TWL and MWL (Kopardekar and Magyarits, 2003) poses a problem to understand other subtler components of workload. Thus, experimental designs should attempt to control for the effect of traffic density in order to focus on the isolated effects of other factors.

## Author statement

**Jorge Ibáñez-Gijón:** Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Visualization; Writing - original draft; Writing - review & editing, **David Travieso:** Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing - original draft; Writing - review & editing, **Jose A. Navia:** Data curation; Investigation; Methodology; Visualization; Writing - review & editing, **Aitor Montes** Data curation; Formal analysis; Software; Validation; Visualization; **David M. Jacobs:** Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; **Patricia L. Frutos**: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision.

## Declaration of competing interest

We have no known conflict of interest to disclose.

## References


Abildstrom, S.Z., Jensen, B.T., Agner, E., Torp-Pedersen, C., Nyvad, O., Wachtell, K., BEAT Study Group, 2003. Heart rate versus heart rate variability in risk prediction after myocardial infarction. J. Cardiovasc. Electrophysiol. 14 (2), 168–173. https://doi.org/10.1046/j.1540-8167.2003.02367.x.

Aricò, P., Borghini, G., Di Flumeri, G., Bonelli, S., Golfetti, A., Graziani, I., et al., 2017. Human factors and neurophysiological metrics in air traffic control: a critical review. IEEE Rev. Biomed Eng. 10, 250–263. https://doi.org/10.1109/RBME.2017.2694142.

Bingham, G.P., 1988. Task-specific devices and the perceptual bottleneck. Hum. Mov. Sci. 7 (2–4), 225–264.

Boag, C., Neal, A., Loft, S., Halford, G.S., 2006. An analysis of relational complexity in an air traffic control conflict detection task. Ergonomics 49 (14), 1508–1526. https://doi.org/10.1080/00140130600779744.

Boucsein, W., Fowles, D.C., Grimnes, S., Ben-Shakhar, G., Roth, W.T., et al., 2012. Publication recommendations for electrodermal measurements. Psychophysiology 49 (8), 1017–1034. https://doi.org/10.1111/j.1469-8986.1981.tb03024.x.

Brookings, J.B., Wilson, G.F., Swain, C.R., 1996. Psychophysiological responses to changes in workload during simulated air traffic control. Biol. Psychol. 42 (3), 361–377. https://doi.org/10.1016/0301-0511(95)05167-8.

Cain, B., 2007. A Review of the Mental Workload Literature. Defense Research and Development, Toronto (Canada).

Charles, R.L., Nixon, J., 2019. Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. 74, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028.

Chatterji, G., Sridhar, B., 2001. Measures for air traffic controller workload prediction. In: 1st AIAA, Aircraft, Technology Integration, and Operations Forum, p. 5242. https://doi.org/10.2514/6.2001-5242.

Colle, H.A., Reid, G.B., 1998. Context effects in subjective mental workload ratings. Hum. Factors 40 (4), 591–600. https://doi.org/10.1518/2F001872098779649283.

Di Stasi, L.L., Marchitto, M., Antolí, A., Baccino, T., Cañas, J.J., 2010. Approximation of on-line mental workload index in ATC simulated multitasks. J. Air Transport. Manag. 16 (6), 330–333. https://doi.org/10.1016/j.jairtraman.2010.02.004.

Djokic, J., Lorenz, B., Fricke, H., 2010. Air traffic control complexity as workload driver. Transport. Res. C Emerg. Technol. 18 (6), 930–936. https://doi.org/10.1016/j.trc.2010.03.005.

Durso, F.T., Manning, C.A., 2008. Air traffic control. Reviews of Human Factors and Ergonomics 4, 195–244. https://doi.org/10.1518/155723408X342853.

Edwards, T., Homola, J., Mercer, J., Claudatos, L., 2017. Multifactor interactions and the air traffic controller: the interaction of situation awareness and workload in association with automation. Cognit. Technol. Work 19 (4), 687–698. https://doi.org/10.1007/s10111-017-0445-z.

Endsley, M.R., 1995. Measurement of situation awareness in dynamic systems. Hum. Factors 37 (1), 65–84. https://doi.org/10.1518/001872095779049499.

Endsley, M.R., Smolensky, M.W., 1998. Situation awareness in air traffic control: the picture. In: Smolensky, M.W., Stein, E.S. (Eds.), Human Factors in Air Traffic Control. Academic Press, pp. 115–154.

EU, 2014. Establishment of the Pilot Common Project Supporting the ATM Master Plan. OJ, 28.6.2014). Regulation 716/2014.

EUROCONTROL, 2004. Cognitive Complexity in Air Traffic Control a Literature Review. EEC, Brussels note no. 04/04.

Falkland, E.C., Wiggins, M.W., 2019. Cross-task cue utilization and situational awareness in simulated air traffic control. Appl. Ergon. 74, 24–30. https://doi.org/10.1016/j.apergo.2018.07.015.

Fothergill, S., Loft, S., Neal, A., 2009. ATC-lab Advanced: an air traffic control simulator with realism and control. Behav. Res. Methods 41 (1), 118–127. https://doi.org/10.3758/BRM.41.1.118.

Frutos, P.L.D., Rodríguez, R.R., Zhang, D.Z., Zheng, S., Cañas, J.J., Muñoz-de-Escalona, E., 2019. 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, pp. 85–104.

Fürstenau, N., Radüntz, T., 2022. Power law model for subjective mental workload and validation through air traffic control human-in-the-loop simulation. Cognit. Technol. Work 24 (2), 291–315. https://doi.org/10.1007/s10111-021-00681-0.

Gopher, D., Braune, R., 1984. On the psychophysics of workload: why bother with subjective measures? Hum. Factors 26 (5), 519–532. https://doi.org/10.1177/2F001872088402600504.

Greco, A., Valenza, G., Lanata, A., Scilingo, E.P., Citi, L., 2015. cvxEDA: a convex optimization approach to electrodermal activity processing. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (4), 797–804. https://doi.org/10.1109/TBME.2015.2474131.

Hart, S.G., 2006. NASA-task load index (NASA-TLX); 20 years later. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50 (9), 904–908. https://doi.org/10.1177/154193120605000909. Sage CA: Los Angeles, CA: Sage publications.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9. North-Holland.

Hilburn, B., 2004. Cognitive complexity in air traffic control: a literature review. EEC note 4, 1–80, 04.

Histon, J.M., Hansman, R.J., 2008. Mitigating Complexity in Air Traffic Control: the Role of Structure-Based Abstractions. Department of Aeronautics and Astronautics, Massachusetts Institute of Technology.

Hopkin, V.D., 2017. Human Factors in Air Traffic Control. CRC Press, London.

Huey, B.M., Wickens, C.D., 1993. Workload Transition: Implications for Individual and Team Performance. National Research Council, Washington DC.

Hughes, A.M., Hancock, G.M., Marlow, S.L., Stowers, K., Salas, E., 2019. Cardiac measures of cognitive workload: a meta-analysis. Hum. Factors 61 (3), 393–414. https://doi.org/10.1177/0018720819830553.

Ibáñez-Gijón, J., Buekers, M., Morice, A., Rao, G., Mascret, N., Laurin, J., Montagne, G., 2017. A scale-based approach to interdisciplinary research and expertise in sports. J. Sports Sci. 35 (3), 290–301. https://doi.org/10.1080/02640414.2016.1164330.

ICAO International Civil Aviation Organization, 2007. Global Air Transport Outlook to 2030. Montreal, Canada.

Jordan, C.S., Brennen, S.D., 1992. Instantaneous Self-Assessment of Workload Technique (ISA). Defense Research Agency, Portsmouth.

Kopardekar, P., Magyarits, S., 2003. Measurement and prediction of dynamic density. In Proceedings of the 5th USA/EUROPE air traffic management R&D Seminar 139.

Kostenko, A., Rauffet, P., Chauvin, C., Coppin, G., 2016. A dynamic closed-looped and multidimensional model for Mental Workload evaluation. IFAC-PapersOnLine 49 (19), 549–554. https://doi.org/10.1016/j.ifacol.2016.10.621.

Loft, S., Sanderson, P., Neal, A., Mooij, M., 2007. Modeling and predicting mental workload in en route air traffic control: critical review and broader implications. Hum. Factors 49 (3), 376–399. https://doi.org/10.1518/001872007X197017.

Mach, S., Storozynski, P., Halama, J., Krems, J.F., 2022. Assessing mental workload with wearable devices–Reliability and applicability of heart rate and motion measurements. Appl. Ergon. 105, 103855 https://doi.org/10.1016/j.apergo.2022.103855.

Majumdar, A., Ochieng, W.Y., McAuley, G., Lenzi, J.M., Lepadatu, C., 2004. The factors affecting airspace capacity in Europe: a cross-sectional time-series analysis using simulated controller workload data. J. Navig. 57 (3), 385–405. https://doi.org/10.1017/S0373463304002863.

Marchitto, M., Benedetto, S., Baccino, T., Cañas, J.J., 2016. Air traffic control: ocular metrics reflect cognitive complexity. Int. J. Ind. Ergon. 54, 120–130. https://doi.org/10.1016/j.ergon.2016.05.010.

Matsumoto, H., 2007. International air network structures and air traffic density of world cities. Transport. Res. E Logist. Transport. Rev. 43 (3), 269–282. https://doi.org/10.1016/j.tre.2006.10.007.

Metzger, U., Parasuraman, R., 2001. The role of the air traffic controller in future air traffic management: an empirical study of active control versus passive monitoring. Hum. Factors 43 (4), 519–528. https://doi.org/10.1518/001872001775870421.

Metzger, U., Parasuraman, R., 2017. Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. In: Decision Making in Aviation 345–360 (Routledge).

Mitchell, D.K., 2000. Mental Workload and ARL Workload Modeling Tools. Army Research Lab Aberdeen Proving Ground MD.

Mohammed, G., El Bekkaye, M., 2021. Fuzzy Dynamic Airspace Sectorization Problem. In Machine Intelligence and Data Analytics for Sustainable Future Smart Cities. Springer, Cham, pp. 229–250.

Nixon, J., Charles, R., 2017. Understanding the human performance envelope using electrophysiological measures from wearable technology. Cognit. Technol. Work 19 (4), 655–666. https://doi.org/10.1007/s10111-017-0431-5.

Pagnotta, M., Jacobs, D.M., de Frutos, P.L., Rodríguez, R., Ibáñez-Gijón, J., Travieso, D., 2021. Task difficulty and physiological measures of mental workload in air traffic

control: a scoping review. Ergonomics 65 (8), 1095–1118. https://doi.org/10.1080/00140139.2021.2016998.

Radüntz, T., Mühlhausen, T., Freyer, M., Fürstenau, N., Meffert, B., 2021. Cardiovascular biomarkers' inherent timescales in mental workload assessment during simulated air traffic control tasks. Appl. Psychophysiol. Biofeedback 46 (1), 43–59. https://doi.org/10.1007/s10484-020-09490-z.

Reid, G.B., 1997. Applications using formal measurement theory. Int. J. Cognit. Ergon. 1 (4), 303–313.

Shou, G., Ding, L., 2013. Frontal theta EEG dynamics in a real-world air traffic control task. Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 5594–5597. https://doi.org/10.1109/EMBC.2013.6610818.

Socha, V., Hanáková, L., Valenta, V., Socha, L., Ábela, R., Kušmírek, S., et al., 2020. Workload assessment of air traffic controllers. Transport. Res. Procedia 51, 243–251. https://doi.org/10.1016/j.trpro.2020.11.027.

Sperandio, J.C., 1971. Variation of operator's strategies and regulating effects on workload. Ergonomics 14 (5), 571–577. https://doi.org/10.1080/00140137108931277.

Suárez, M.Z., Valdés, R.M.A., Moreno, F.P., Jurado, R.D.A., de Frutos, P.M.L., Comendador, V.F.G., 2022. How much workload is workload? A human neurophysiological and affective-cognitive performance measurement methodology for ATCOs. Aircraft Eng. Aero. Technol. 94 (9), 1525–1536. https://doi.org/10.1108/AEAT-11-2021-0328.

Subramanian, S., Barbieri, R., Brown, E.N., 2019. A Systematic Method for Preprocessing and Analyzing Electrodermal Activity. , July. In 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6902–6905. https://doi.org/10.1109/EMBC.2019.8857757.

Tattersall, A.J., Foord, P.S., 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics 39 (5), 740–748. https://doi.org/10.1080/00140139608964495.

The jamovi project, 2022. jamovi. Version 2.3) [Computer Software]. Retrieved from. https://www.jamovi.org.

Tobaruela, G., Schuster, W., Majumdar, A., Ochieng, W.Y., Martinez, L., Hendrickx, P., 2014. A method to estimate air traffic controller mental workload based on traffic clearances. J. Air Transport. Manag. 39, 59–71. https://doi.org/10.1016/j.jairtraman.2014.04.002.

Vollmer, M., 2015. A Robust, Simple and Reliable Measure of Heart Rate Variability Using Relative RR Intervals. In Computing in Cardiology Conference (CinC). IEEE. https://doi.org/10.1109/2Fcic.2015.7410984.

Wee, H.J., Lye, S.W., Pinheiro, J.-P., 2018. A spatial, temporal complexity metric for tactical air traffic control. J. Navig. 71 (5), 1040–1054. https://doi.org/10.1017/S0373463318000255.

Wickens, C.D., 2008. Multiple Resources and mental workload. Hum. Factors 50 (3), 449–455. https://doi.org/10.1518/001872008X288394.

Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17. https://doi.org/10.1080/00140139.2014.956151.