Applied Ergonomics 115 (2024) 104162

Contents lists available at ScienceDirect

Applied Ergonomics

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

The effects of increased mental workload of air traffic controllers on time perception: Behavioral and physiological evidence

Eirini Balta, Andreas Psarrakis, Argiro Vatakis *

Multisensory and Temporal Processing Lab (MultiTimeLab), Department of Psychology, Panteion University of Social and Political Sciences, Athens, Greece

ARTICLE INFO

Keywords:
Mental workload
Time perception
Air traffic controllers
Timing
Physiological measurements
Passage of time
Time estimation

ABSTRACT

Research has shown that timing is modulated by mental workload, making duration judgments a measure of cognitive demand, alongside subjective assessments, and physiological measurements. Yet, it is unclear whether such findings can be extended in less controlled setups. By employing air traffic controllers in a real aviation environment, we tested whether tasks with different levels of cognitive load can affect their timing behavior. Participants completed temporal production, verbal estimation, and passage of time judgments, while actively engaging in real flight control sessions. Subjective assessments of task demands, as well as physiological responses (cardiac and electrodermal activity) were also measured. Accuracy of the produced intervals was measured at two distinct phases of the flight (during low-load cruising vs. high-load landing) and under two different task load manipulations (controlling one vs. two helicopters and speaking in native vs. non-native language). Analysis of interval production accuracy showed that during the high-load landing phase significant overproductions were made, compared to the low-load cruising phase, and landing two helicopters led to greater overproductions compared to landing only one. The duration of the two-helicopter sessions was significantly overestimated compared to the single-helicopter ones, and the passage of time was felt significantly faster. Subjective assessments of workload were positively correlated with the temporal estimations and passage of time judgments, and skin responses were positively correlated with the produced intervals. Overall, our results are consistent with past research, suggesting that mental workload modulates time perception in complex, real-world environments, thus making timing behavior a reliable index of the workload changes.

1. Introduction

Mental workload is one of the most prominent concepts in Ergonomics, when one considers the optimization of operators' performance in dynamically changing work environments (Young et al., 2015). Widely accepted as a multidimensional concept, mental workload (or simply workload) has been examined by many theories and defined in different ways, which selectively focus on the multiple facets comprising workload (for an extensive review see Longo et al., 2022). By a comparative semantic analysis of the different definitions of workload, Longo et al. identified the two basic parameters, the task and the operator, as well as their interaction, as the source of workload. Consequently, mental workload can be considered as the cognitive load that a specific task imposes on the operator who possesses a finite source of cognitive resources.

The effects of mental workload on performance are of particular importance in complex environments such as those of air traffic control. Air traffic controllers usually handle tasks with increased complexity and versatility, which, in turn, increase the cognitive demands needed to perform at high levels. In their study, Pape et al. (2001) found that air traffic control-related incidents or accidents are rare, but when they do happen they are most likely to be attributed to human error. Furthermore, they found that these incidents were caused mainly by attentional failures and memory lapses (Pape et al., 2001). Thus, the effect of workload becomes apparent and the need to identify sudden changes of load are of high priority, especially in dynamically changing working environments.

Mental workload in air traffic control is affected by the complexity of the tasks the controllers must perform (Hilburn, 2004). Amongst the factors that render a task more difficult is the number of aircrafts under control (Edwards et al., 2017; Hurst and Rose, 1978; Stein, 1985). Hilburn (2004) notes that the positive association between the number of controlled air vehicles and perceived levels of workload by the air traffic controllers is the one most frequently found, thus, making the number of

* Corresponding author. Department of Psychology, Panteion University of Social and Political Sciences, 136 Syngrou Ave., 17671 Athens, Greece.
E-mail address: argiro.vatakis@gmail.com (A. Vatakis).

https://doi.org/10.1016/j.apergo.2023.104162
Received 19 May 2023; Received in revised form 27 October 2023; Accepted 30 October 2023
Available online 4 November 2023
0003-6870/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

controlled aircrafts a robust index of workload. Other factors that render a task more complex and can increase perceived load are the number of altitude transitions (Cardosi and Murphy, 1995), the variations of flight directions and the frequency of traffic problems (Hilburn, 2004), the weather conditions (Kontogiannis and Malakis, 2017; Mogford et al., 1994), and the language of communication (Estival et al., 2016; Hopkin, 1982; Lin, 2021). This latter factor is of particular importance, especially when considering that linguistic difficulties and the use of inadequate language are viewed as contributing factors to the deadliest accident in aviation history (Joint Report: KLM-PAA, 1978; Roitsch et al., 1977). Since the English language is the de facto international language of aviation, air traffic controllers may experience higher levels of workload, either because they are non-native speakers of Aviation English (Hasegawa et al., 2002; Wu et al., 2020), or because they communicate with non-native English-speaking pilots (Tiewtrakul and Fletcher, 2010). Furthermore, because speaking in a non-native language both affects and is affected by the overall task load (Farris et al., 2008; Prinzo et al., 2010), defining and measuring how workload changes under complex situations becomes more challenging.

Ways to measure mental workload have been driven by the need to decrease human errors in critical working environments, like aviation, military, emergency services (Chen et al., 2016), but also in learning conditions, in order to improve learning performance (Paas et al., 2003). Workload measuring techniques can be distinguished in performance-based techniques -in the primary and a secondary task-, subjective procedures, and physiological measurements. Primary task performance-based techniques rely on the hypothesis that as the levels of workload increase, performance in the task is expected to decrease (Tsang and Vidulich, 2006). Nevertheless, it has also been suggested that in cases of small workload changes, this association may not be valid, since performance can be retained by assigning available cognitive resources, while in significant workload changes, individual strategies and decision making choices may be employed in order to maintain load in acceptable levels, while choosing specific tasks to underperform (O'Donnell and Eggemeier, 1986). It is, thus, suggested that primary task performance measures do not always evaluate the changes in workload (Hart and Wickens, 1990). On the other hand, secondary task performance techniques rely on the evaluation of performance of an additional task, which is executed with the processing resources that remain after allocating what is needed for performing the primary task. Consequently, secondary task performance decreases as cognitive demands in the primary task increase, thus indirectly reflecting changes in workload. The most serious limitation in using a secondary task as a load index is that it may interfere with the primary task (Meshkati and Loewenthal, 1988). A secondary task that has been proposed as an appropriate measure of workload is a time perception task (Hart et al., 1978). Duration judgment tasks, whether time estimation or temporal production, have been found to be sensitive to changes in workload (Zakay and Shub, 1998), while at the same time unobtrusive to the primary task performance (Brown, 1997, 2006). In cases of increased mental workload, prospective temporal estimations were shortened and temporal productions lengthened, whereas in retrospective timing tasks the reverse effects were obtained (Baldauf et al., 2009; Block et al., 2010; Hart et al., 1978; Liu and Wickens, 1994; Zakay and Shub, 1998), thus making performance in a timing task a suitable index of workload.

Most studies that have examined the use of a timing task for measuring changes in workload have been carried out in strictly controlled environments, using artificial tasks and simple stimuli. This poses the question whether basic research findings on timing can be applied on real-world environments in an ecologically valid way (Van Rijn, 2018). Though some studies that used more complex setups, like flight simulators (Zakay and Shub, 1998) and gaming centers (Tobin et al., 2010), or more complex, event-based stimuli (Schlichting et al., 2018), have replicated basic findings in the timing literature, other studies have not shown such effects. That is, studies have shown, for example, that: a) duration production can be a measure of workload (Zakay and Shub, 1998), b) prospective timing is overestimated compared to retrospective timing and shorter durations are overestimated more than longer intervals (Tobin et al., 2010), c) the scalar property and the temporal context effect also apply to naturalistic stimulation (Schlichting et al., 2018). On the other hand, by using naturalistic events, like knocking a door, or eating an apple, in the auditory, visual, or audiovisual modality, Boltz (2005) found no modality difference in the reproduction times of the scenes, contrary to the expectation of an overestimation of the scenes in the auditory or audiovisual modality as compared to the visual ones (Wearden et al., 1998). More recently, Riemer et al. (2021) found that naturalistic visual scenes differentially affect the time precision of younger and older adults, with the timing performance of the latter being lower when the to-be-timed stimuli were part of more complex scenes than when they were isolated. Similarly, Tachmatzidou and Vatakis (2023) found that an unexpected stimulus violating the semantic coherence of a naturalistic scene was not overestimated, as expected by the temporal oddball effect (Tse, 2004), and that this effect was dependent on the manipulation of attention. Such challenges of the ecologic validity of basic time perception findings when moving to real-world settings points to the need of further investigating the use of a timing task to measure mental workload.

On the other hand, subjective measurements are based on the person's judgment on the load imposed by the task, and have been tested as a means of workload assessment in diverse domains, like air traffic control, automobile driving, medical profession, and use of computers and portable technology (Hart, 2006). Though subjective judgments of workload have been criticized as not being able to detect cognitive processes that are affected by increases in the load (O'Donnell and Eggemeier, 1986), they nevertheless have been proven to possess a global sensitivity to the variation of those factors that affect workload and, thus, the demand for increased processing resources (Wierwille and Eggemeier, 1993). Especially, the National Aeronautics and Space Administration Task Load Index (NASA-TLX), is a well-established assessment technique, based on a multi-dimensional questionnaire, used in diverse environments (Hart, 2006), with increased sensitivity even in low levels of workload (Rubio et al., 2004), thus being particularly appropriate for evaluating the load conditions in a real-world environment.

Similarly, changes in the level of task load have been found to affect the physiological states, causing heart, electrodermal, respiratory, ocular, and brain activity changes (Charles and Nixon, 2019). Heart rate variability has been found to decrease with increasing load (Paas et al., 1994; Van Roon et al., 2004), electrodermal activity increases (Nourbakhsh et al., 2017; Shi et al., 2007), pupil diameter, blink, and fixation frequency increase (Van Orden et al., 2001), breathing rate increases (Grassmann et al., 2016), and electroencephalography signals change (Antonenko et al., 2010). In a study examining 33 experiments that used physiological measures to assess workload, Ayers et al. (2021) found that nearly all the physiological measures that were employed (i.e., heart, respiration, eye, skin, brain measures) can at some level detect workload changes, with a varying degree of sensitivity that may depend on the nature of the task (e.g., electrodermal activity might be more sensitive in tasks that cause abrupt load changes). In the simulation tasks studied (motorcar driving, engineering skills, military exercises, and surgery), electrodermal activity was more sensitive (though only one study was examined), and cardiac activity less sensitive in measuring workload. With recent technological advances, physiological measurements become more readily available, in real-time conditions, by using wearable devices of relatively low cost (Gjoreski et al., 2018; Jaiswal et al., 2021; Romine et al., 2020). As a result, physiological measurements can be used in combination with performance metrics and subjective judgments, to detect workload changes in the most effective way.

The current study tested the effect of mental workload on time perception in a real-world scenario, by employing the dual-task paradigm in an air traffic control setup. We examined whether professional

air traffic controllers performing a real flight control task with varying degrees of difficulty would exhibit changes in their timing performance on a parallel production and a consequent time estimation and passage of time judgment task. We hypothesized that any changes in performance in the timing task will be induced by the differential level of workload of the non-timing task. Specifically, we hypothesized that increasing the number of helicopters that the controller simultaneously maintains, changing the language of communication from native to non-native, and differentiating between the cruising and landing phase of the flight would lead to overproductions, time overestimation, and faster passage of time judgments, due to the increase in the workload. The current study aims to extend the existing evidence on time perception in more naturalistic situations, where dynamic sequences of events and stimuli are continuously monitored and timed. We examined whether the timing performance, along with the subjective assessments of the task demands and the physiological measurements of the controllers would be an index for the mental workload changes, and we discuss possible implications of time estimation accuracy in operations like that of air traffic control.

2. Method

2.1. Participants

Twelve air traffic controllers (4 females), aged between 27 and 47 years ($M = 36.86$ $SD = 6.6$), participated in the experiment. The participants were all professionals, certified by the Civil Aviation Authority of the Ministry of Infrastructure and Transport, employed in a Greek military airport, and with 3–23 years of professional experience ($M = 11.83$, $SD = 6.71$). Their native language was Greek, they all had an excellent knowledge of English as a second language, and they had undergone the necessary training in air traffic management terminology. All participants were naïve as to the purpose of the experiment. The experiment was performed in accordance with the ethical standards laid down in the 2013 Declaration of Helsinki, the ethical approval of the study by the University's ethics committee (19/15-5-2022), and informed consent was obtained from all participants.

2.2. Apparatus

The experiment was conducted in the air traffic control premises located in the military airport. The main equipment used was the communication device, and the digital recorder of the control tower. The communication device was an Integrated Communications (ICOM) VHF air-band base station radio, equipped with a hand-held microphone that the participants used to communicate with the pilots. The digital recorder's press key was used for the production timing tasks. Furthermore, the EmotiBit open-source wearable device (www.emotibit.com) was used to measure the cardiac and electrodermal activity throughout experimentation. The 3-wavelength photoplethysmogram (PPG) sensor was used to monitor blood volume changes at a sampling rate of 25 Hz, and the electrodermal activity (EDA) sensor was used to monitor skin conductance at a sampling rate of 15 Hz. The physiological data were recorded on a secure digital card (SD card) for offline processing.

2.3. Design

In all experimental conditions, the participant, wearing the EmotiBit device, completed a flying control session, while performing a timing production task, a time estimation task, and a passage of time judgment. Each session took place during scheduled training flights for the pilots of the military helicopters, where the pilots had to take-off, perform a full circle over the airport, and land. During each session, the air traffic controller communicated with the pilot of the helicopter, and provided all the necessary information for performing the flight. Specifically, during the take-off and landing phases, the controller gave clearance for the helicopter's take-off or landing, informed the pilot on the airstrip to use, and gave information on the wind direction and speed, and on the barometric pressure. During the cruising phase of the flight, the air traffic controller maintained visual contact with the helicopter, while it circled the airport flying at a constant speed. No other air traffic was present over the controlled airport space, and no unexpected communication between the air traffic controller and the pilot was recorded at any point, beyond the one reported and foreseen. This design was intentional, in order to control for any confounding variables that might differentially affected the performance of the air traffic controllers in the timing tasks across conditions and to adhere to the strict safety protocol of military flights.

During the flying control task, each air traffic controller completed two timing production tasks, each at two distinct stages of the flight. Specifically, each participant was instructed to produce an interval by pressing a key using the control tower equipment. The production times ranged randomly from 25 to 35 s.[1] The first time production task was performed while the helicopter was cruising over the airport, approximately 15 s after its take-off. The second production task was performed during the landing of the helicopter. Each stage of the flight posed different demands on the air traffic controller. During the cruising phase the controller only maintained visual contact with the helicopter, while during the landing phase the controller was actively engaged in the procedure. Thus, the two production tasks were performed under different workload conditions, the low-level cruising production, and the high-level load landing production.

At the end of each experimental session, each participant made a verbal estimation of the total duration of the session. Additionally, the participants gave a rating on how slow or fast they felt the time had passed during the session (i.e., passage of time judgment). Lastly, they completed the NASA-TLX questionnaire, giving their subjective assessment on the task workload demands. Fig. 1 shows an overview of the experimental design.

Overall, four different types of experimental sessions were designed according to the workload demands of the task. The level of workload was manipulated based on the number of helicopters controlled by the air traffic controller, and the language of communication between the controller and the pilot of the helicopter. In the low-level load condition, one helicopter was under the participant's control, while in the high-level load condition, two helicopters were simultaneously controlled. Also, regarding the language of communication, in the low-level condition the controller was speaking to the pilot in their native language (i.e., Greek), while in the high-level condition, the controller was communicating with the pilot in a non-native language (i.e., English). The two manipulations (number of helicopters and language of communication) were fully crossed, thus creating the four experimental sessions: one helicopter-Greek language (low-low), one helicopter-English language (low-high), two helicopters-Greek language (high-low), two helicopters-English language (high-high). An additional in-session condition, for the time production task only, was created by the different flying phases of the helicopter (i.e., cruising, landing) during which the interval production was made.

The experiment took place in two consecutive days, between the hours of 08:00 and 14:00. The days were selected after considering the weather forecast provided by the Hellenic National Meteorological Service and ensuring that the weather conditions would be as similar as possible to control for the confounding effect of weather on the

[1] These timings were selected given that according to Hart et al. (1978), temporal intervals ranging from 1 to 30 s are suitable for timing production tasks given that during these periods the attentional demands of the primary air traffic control task are well reflected. Furthermore, according to Block et al. (2010), the estimation of a target duration less than 60 s relies less on long-term memory processes, and, in previous studies on the effect of mental workload on timing, target durations varied between 20 and 30 s (Block et al., 2010).

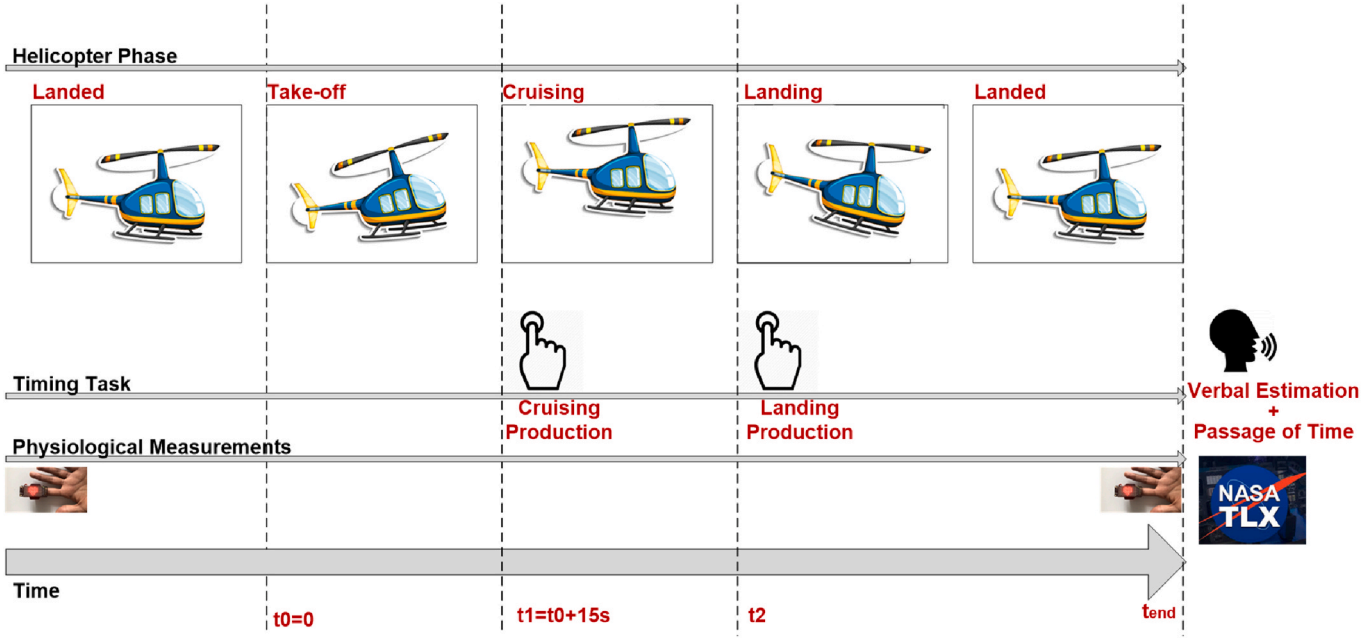

Fig. 1. Schematic Representation of the Experimental Session Design. *Note.* Each air traffic controller completed two parallel tasks per session. During the flight control task, the participant controlled a helicopter while it performed a full circular flight over the airport area. At two distinct stages of the flight (during cruising and landing), the controller performed a time production task, and at the end of the session gave a verbal estimation of the duration of the session, made a passage of time judgment, and completed the NASA-TLX questionnaires. During the whole session, the cardiac and electrodermal activity of the controller were continuously monitored by the EmotiBit wearable device.

experimental results. Each participant participated in two sessions per day, for a total of four sessions. The order of the sessions was randomized. Each session lasted approximately 3 min, which was the total time taken by the helicopter to take off, circle over the airport, and land. The start of the session was defined as the time the helicopter (or the first helicopter in the case of two helicopters) began taking off, and the end of the session was the time the landing of the helicopter (or second helicopter) was concluded.

2.4. Procedure

Prior to the start of each session the participants wore the EmotiBit device and received detailed instructions for the experimental procedure. They were informed that they will be asked to perform several time production tasks by pressing a key on the digital recorder using their dominant hand, while performing the flight control task. They were asked not to follow any counting strategies during the production task, and all timing devices were removed from their sight. The EmotiBit wearable was attached to the index finger of the participants' non-dominant hand, and they were instructed to keep this hand as still as possible to avoid any motion artifacts.

After the start of the flight task, and while the helicopter was in cruising phase, the experimenter verbally asked the participants to produce a specified interval, and the participants pressed the assigned key for the appropriate time. The second production instruction was given to the participants while the helicopter was in the landing phase. At the end of each session, the participants were asked to verbally estimate how long the session lasted. They were also asked to answer the question "How fast or slow did you feel the time passed during this session?", giving their answer using a 5-point Likert scale (1-very slow, 5-very fast). Finally, they filled out the NASA-TLX questionnaire, regarding the task's workload.

3. Results

3.1. Behavioral data analysis

Two measures of timing performance accuracy, and one measure of passage of time judgment were analyzed as to how they were affected by the different workload demands in each experimental session and in-session condition. The accuracy of the production timings was defined as the ratio of the produced interval to the verbally requested one, while the accuracy of the estimation timings was defined as the verbally estimated interval to the actual duration of the session. Accuracy values greater that one were equal to overproduction and overestimation of the intervals, while accuracy values less than one were equal to underproduction and underestimation of time. For the passage of time analysis, the actual ratings given by the participants were used.

3.1.1. Accuracy of production timings

A three-way repeated measures analysis of variance (ANOVA) was conducted with three within-subjects factors: i) the number of helicopters simultaneously controlled by the air traffic controller (Helicopters, with 2 levels: one vs. two), ii) the language of communication between the controller and the pilot (Language, with 2 levels: Greek vs. English), and iii) the flight phase during which the interval production was made (Phase, with 2 levels: cruising vs. landing). The accuracy of production timings was the dependent variable. Outlier detection was based on Rosner's Generalized Extreme Studentized deviate test (Rosner, 1983) to account for the sample size. Robust ANOVA analysis was conducted using the WRS R-package (Wilcox and Schönbrodt, 2014) to account for data deviations from normality. Bonferroni corrections were applied for all pairwise comparisons.

The analysis showed a significant main effect of the number of Helicopters [$F(1,11) = 35.093$, $p < .0001$, $\eta^2 = 0.216$], a significant main effect of the flight Phase [$F(1,11) = 60.118$, $p < .0001$, $\eta^2 = 0.276$], and a significant interaction between Helicopters and Phase [$F(1,11) = 34.624$, $p = .000106$, $\eta^2 = 0.126$] (see Fig. 2). Pairwise comparisons for repeated measures ANOVA showed that the numbers of helicopters

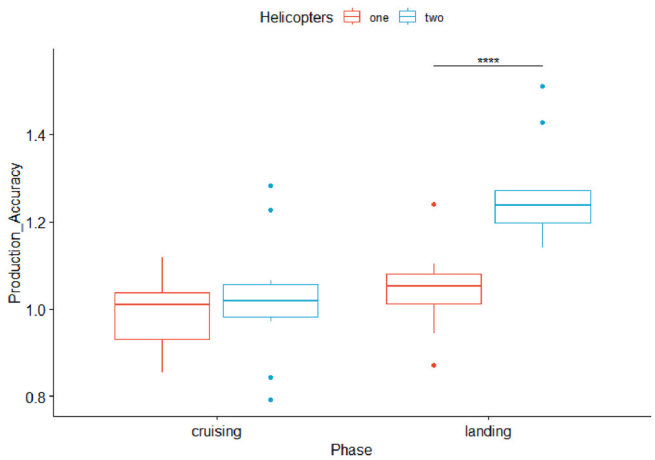

Fig. 2. Effect of Helicopters Factor at each Phase Condition on Production Accuracy. *Note.* Mean Production Accuracy for the different number of helicopters and flight phase condition. Significant differences are indicated by the asterisks. The error bars represent the standard error of the mean.

affected the production timings during the landing phase. Specifically, when the controllers performed the time production task while the helicopters were in the landing phase, mean production ratio (time produced to time requested) was greater in the case of two Helicopters ($M_{\text{two-landing}} = 1.26$, $SD_{\text{two-landing}} = 0.135$) as compared to one Helicopter ($M_{\text{one-landing}} = 1.04$, $SD_{\text{one-landing}} = 0.112$), $t(23) = 7.33$, $p < .0001$. This effect was not observed when either one or two helicopters were in the cruising phase ($M_{\text{two-cruising}} = 1.02$, $SD_{\text{two-cruising}} = 0.143$, $M_{\text{one-cruising}} = 0.989$, $SD_{\text{one-cruising}} = 0.089$), $t(23) = 1.67$, $p = .108$.

Additionally, the flight phase affected the production timings when either one or two helicopters were simultaneously controlled (see Fig. 3). The production ratio was significantly greater when the production task was performed during the landing Phase ($M_{\text{two-landing}} = 1.26$, $SD_{\text{two-landing}} = 0.135$) than when the helicopters where in the cruising Phase ($M_{\text{two-cruising}} = 1.02$, $SD_{\text{two-cruising}} = 0.143$), $t(23) = 7.94$, $p < .0001$. The same effect was observed for the case of one controlled helicopter ($M_{\text{one-landing}} = 1.04$, $SD_{\text{one-landing}} = 0.112$, $M_{\text{one-cruising}} = 0.989$, $SD_{\text{one-cruising}} = 0.089$), $t(23) = 2.93$, $p = .007$. The main effect of Phase showed that mean production accuracy was significantly greater in the landing phase compared to the cruising phase ($M_{\text{landing}} = 1.15$, $SD_{\text{landing}} = 0.163$, $M_{\text{cruising}} = 1.01$, $SD_{\text{cruising}} = 0.119$), $t(47) = 6.68$, $p < .0001$. Similarly, the main effect of Helicopters showed that mean

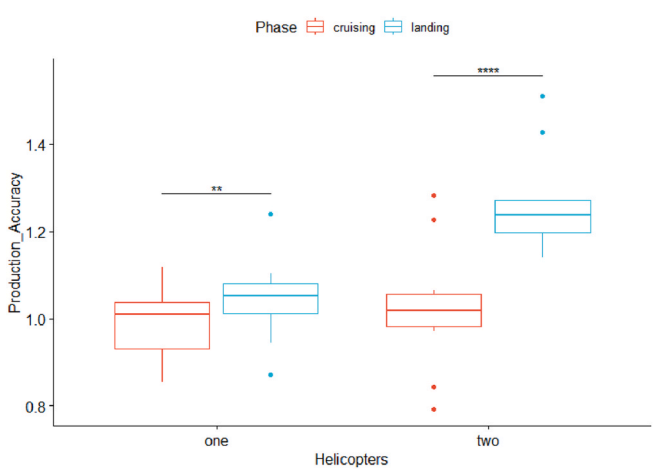

Fig. 3. Effect of Phase Condition at each Helicopter Factor Level on Production Accuracy. *Note.* Mean production accuracy for the different flight phase condition and number of helicopters. Significant differences are indicated by the asterisks. The error bars represent the standard error of the mean.

production accuracy was significantly greater in the case of two helicopters compared to the one-helicopter case ($M_{\text{two}} = 1.14$, $SD_{\text{two}} = 0.182$, $M_{\text{one}} = 1.02$, $SD_{\text{one}} = 0.104$), $t(47) = 5.65$, $p < .0001$.

No interaction effect between the three factors (Helicopters x Language x Phase: $F(1,11) = 1.515$, $p = .244$, $\eta^2 = 0.004$) was obtained. Also, no interaction effect between Helicopters and Language [$F(1,11) = 0.433$, $p = .519$, $\eta^2 = 0.004$], or Language and Phase [$F(1,11) = 0.457$, $p = .513$, $\eta^2 = 0.004$], and no main effect of Language [$F(1,11) = 2.815$, $p = .122$, $\eta^2 = 0.007$] were found.

In order to assess whether the times produced differed from the objective ones (i.e., the timings verbally requested by the experimenter), one-sample t-tests were performed, by comparing the accuracy of production timings to 1 (i.e., perfect accuracy). The analysis showed that for the case of either one or two helicopters in the landing phase, the production accuracy was significantly different from 1 [1 helicopter landing: $t(23) = 1.981$, $p = .02986$, 2 helicopters landing: $t(23) = 9.374$, $p < .0001$]. That is, for both cases, the ratio of produced to requested time was greater than 1 ($M_{\text{one-landing}} = 1.024$, $M_{\text{two-landing}} = 1.259$). In the cruising phase, the production accuracy was not significantly different from 1.

The findings from the production task suggest that manipulating the workload across sessions, by altering the number of helicopters, and within sessions, by distinguishing between the flight phase, affected the accuracy of production timings. As the workload increased, the air traffic controllers made significant overproductions at the high load conditions, compared to the low ones, which are translated as underestimation of time (Brown, 1995; Zakay, 1993; Zakay and Shub, 1998). Notably, in all the cases but the one with the lowest level of workload (one helicopter in the cruising phase), all intervals produced by the air traffic controllers were longer than the ones they were instructed to produce, which equals to a systematic underestimation of time (Block et al., 2010; Hart et al., 1978). Language did not seem to influence the workload as no differences were found between the sessions that used native as compared to non-native language.

3.1.2. Accuracy of estimation timings

A two-way repeated measures ANOVA was conducted to test the effect of: i) the number of helicopters and ii) the language of communication on the accuracy of time estimation. The estimation task was completed by the air traffic controllers once at the end of each session, after the helicopters landed, thus, the phase condition was not a factor in this part of the task. Outlier detection was based on Rosner's Generalized Extreme Studentized deviate and Bonferroni corrections were applied for all pairwise comparisons. Robust ANOVA analysis was conducted to account for data deviations from normality (Wilcox and Schönbrodt, 2014).

The analysis showed a significant main effect of the number of Helicopters on the estimation accuracy [$F(1,11) = 89.4$, $p < .0001$, $\eta^2 = 0.556$]. Post-hoc analysis showed that the mean estimation of the session time when two helicopters were controlled was significantly greater ($M_{\text{two}} = 1.49$, $SD_{\text{two}} = 0.31$) than when only one helicopter ($M_{\text{one}} = 0.837$, $SD_{\text{one}} = 0.309$) had to be supervised by the air traffic controller, $t(23) = 8.28$, $p < .0001$ (see Fig. 4). No effect of Language [$F(1,11) = 1.738$, $p = .214$, $\eta^2 = 0.044$] or an interaction effect between Language and Helicopters [$F(1,11) = 1.171$, $p = .302$, $\eta^2 = 0.025$] were found.

One-sample t-tests compared the difference between the estimated session times and the actual session times. Estimation accuracy was significantly different from 1 for the sessions with one or two helicopters [1 helicopter: $t(23) = -2.578$, $p = .008418$, 2 helicopters: $t(23) = 7.793$, $p < .0001$]. For the case of one helicopter, the ratio of estimated to actual session time was less than 1 ($M_{\text{one}} = 0.837$), whereas in the case of two helicopters the ratio was greater than 1 ($M_{\text{two}} = 1.493$). That is, when controlling one helicopter, the air traffic controllers underestimated the time of the session, whereas when controlling two helicopters the overestimated the duration of the session.

Overall, as workload increased with the number of helicopters, the

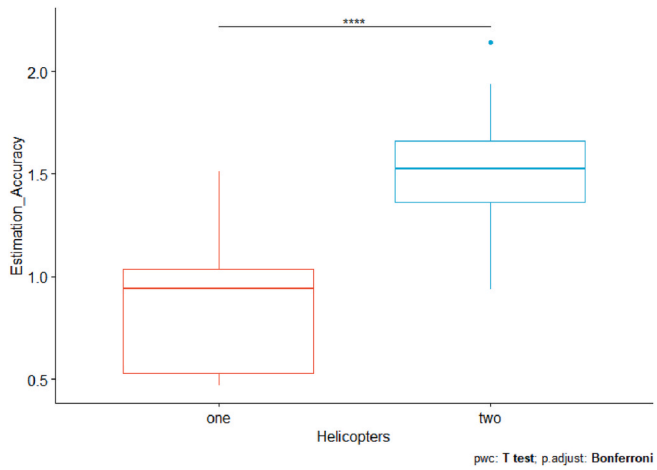

Fig. 4. Mean Estimation Accuracy for the Different Number of Helicopters. *Note.* Mean estimation accuracy for the sessions with one vs. two helicopters. Significant differences are indicated by the asterisks. The error bars represent the standard error of the mean.

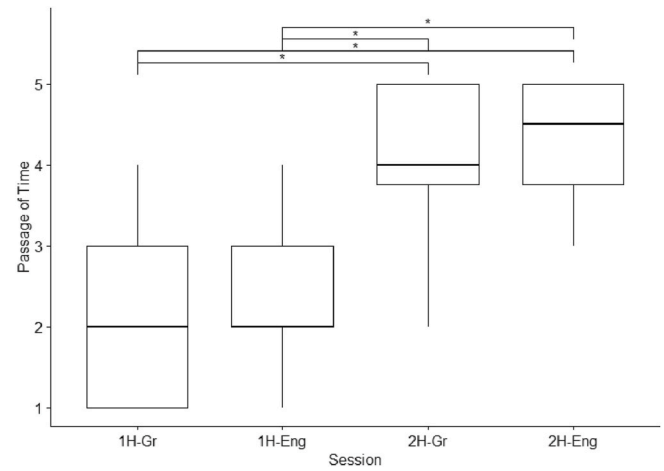

Fig. 6. Mean Passage of Time Ratings across Different Sessions. *Note.* Mean passage of time ratings for the four sessions (1H: one helicopter, 2H: two helicopters, Gr: Greek language, Eng: English language). Higher values in the y-axis mean time felt to pass faster (1: very slow, 5: very fast). Significant differences are indicated by the asterisks. The error bars represent the standard error of the mean.

air traffic controllers made significant duration overestimations of the sessions. Though this result may seem contradicting to the ones previously reported, according to previous literature (Block et al., 2010; Zakay and Fallach, 1984), the immediacy of the temporal task affects its nature, causing a delayed duration judgment to behave more like a retrospective than a prospective one (Block et al., 2010).

3.1.3. Passage of time ratings

For the passage of time judgments (POTJ) that each air traffic controller made after the end of each session, a Friedman test was performed to test for differences in the Likert-scale ratings, between the four different sessions (Helicopters x Language). Fig. 5 shows the distribution of responses for how slow or fast the time felt passing during the whole session, across the different sessions. The analysis showed that the mean POTJ was significantly different at the different sessions, $\chi^2(3) = 29.06$, $p < .0001$, and that the effect of session on the passage of time judgments was large (W = 0.807). Pairwise Wilcoxon signed rank test between the session types showed significant differences between the sessions with one versus two helicopters, irrespective of the language spoken (see Fig. 6). Specifically, time felt to pass significantly faster when two helicopters were controlled than when only one was under the controller's supervision. All pairwise comparisons of the one-helicopter session, either in Greek or English, as compared to the two-helicopter session, in either language, were significant at $p < .05$. The language did not affect the passage of time judgments when the number of controlled helicopters remained the same. Though most often the POTJs have been associated with the hedonic content of the task (Watt, 1991; Wearden, 2005), a study has shown that passage of time judgments are also influenced by workload (Sucala et al., 2011), and time seems to pass faster as task demands increase. Our findings are in accordance with these latter findings, suggesting that in more difficult tasks time seems to "fly".

3.2. Subjective ratings of workload

The subjective rating of workload in each session was based on the NASA TLX questionnaires that the participants completed after the end of each session. These comprise six individual scales of workload, measured on a range of 0–100, that assess different aspects of the task load (e.g., mental demand, temporal demand, effort, frustration).

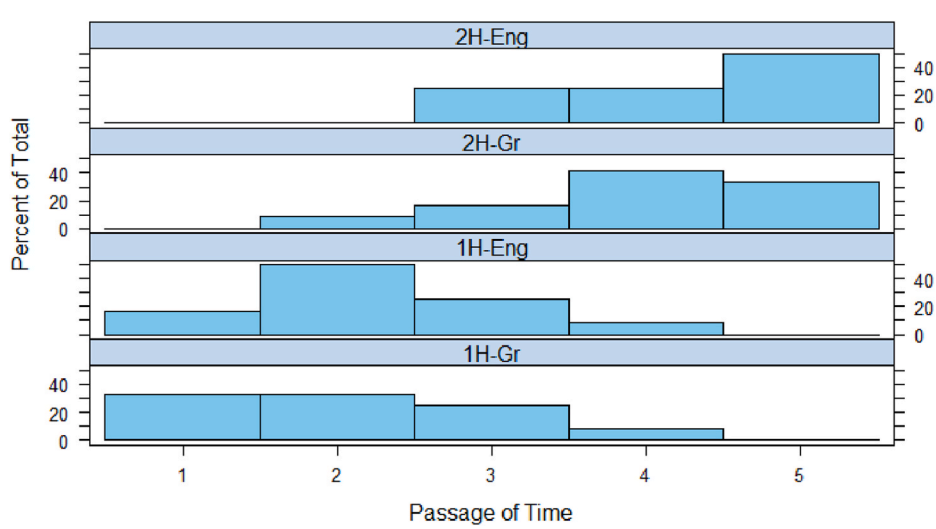

Fig. 5. Frequency of Passage of Time Responses across the Different Sessions. *Note.* The distribution of passage of time responses (1: very slow, 2: slow, 3: normal, 4: fast, 5: very fast) across the four different sessions (1H: one helicopter, 2H: two helicopters, Gr: Greek language, Eng: English language).

Averaging the individual scores results in the raw TLX score for workload.

To assess the degree at which the subjective assessment of load aligned with the actual workload demands in each different session, a two-way repeated measures ANOVA was conducted with the number of helicopters, and the language of communication as factors affecting the raw TLX score. The analysis showed a significant main effect of the number of Helicopters [$F(1,11) = 17.1$, $p = .002$, $\eta^2 = 0.189$] and a significant main effect of Language [$F(1,11) = 24.34$, $p = .000447$, $\eta^2 = 0.031$] on the TLX score. No interaction effect between Helicopters and Language was found [$F(1,11) = 0.31$, $p = .59$, $\eta^2 = 0.00062$]. Post-hoc analysis showed that TLX scores were significantly higher when two helicopters were controlled ($M_{two} = 47.7$, $SD_{two} = 12.5$) as compared to one helicopter ($M_{one} = 38.3$, $SD_{one} = 7.11$), $t(23) = 5.57$, $p < .0001$. Similarly, when controllers were speaking in English, the TLX score was significantly higher ($M_{English} = 44.8$, $SD_{English} = 11.8$) than when speaking in Greek ($M_{Greek} = 41.3$, $SD_{Greek} = 10.3$), $t(23) = 4.43$, $p = .000191$.

To assess the relation between the TLX score and the timing behavior of the air traffic controllers, a correlation analysis was conducted. The results revealed a positive correlation between the Estimation Accuracy and the combined TLX score [$r(46) = 0.3$, $p = .0352$]. Similarly, a positive correlation was found between the Passage of Time Judgment and the TLX score [$r(46) = 0.38$, $p = .00745$] (see Fig. 7).

Overall, the findings suggest that the NASA TXL scores increased with increased session load. The subjective assessments of workload by the air traffic controllers were representative of the increases in task demands. Furthermore, the results showed that there is a correlation between the timing behavior and the subjective ratings across the different sessions, both of which are affected by the workload manipulations.

3.3. Physiological data analysis

Physiological data were collected through the EmotiBit device, which the air traffic controllers wore during the entire flight session. Cardiac activity and electrodermal activity signals were recorded and analyzed to assess whether changes in workload task demands are mirrored in such type of physiological data.

Electrodermal activity (EDA) was recorded during the whole flight session. The EDA signal was processed using the NeuroKit2 Python package (Makowski et al., 2021). The toolkit automatically cleans the signal by removing noise and smoothing the signal, using a low-pass filter with a 3 Hz cutoff frequency and a 4th order Butterworth filter. It then decomposes the signal in two components, the phasic and tonic, and identifies skin conductance responses as the peaks in the phasic component. The analysis focused on the phasic component of the EDA, and the number of skin conductance responses (SCR) was the basic metric. SCR analysis was performed at a per session type and session condition individually. A correlation analysis revealed that the Production Accuracy and the SCRs were positively correlated [$r(94) = 0.21$, $p = .00437$]. As workload increased, temporal intervals were overproduced more, and skin conductance responses increased. Though a three-way repeated measures ANOVA did not reveal any significant effect of workload [$F(1,11) = 0.906$, $p = .362$, $\eta^2 = 0.01$] on the SCR, there was a trend for the EDA activity, as measured by the SCR, to increase in the sessions where English was the language of communication, and the helicopters were in the landing phase as compared to the cruising phase. Fig. 8 shows the mean number of SCR peaks per session type and condition.

Cardiac activity signals were also recorded for the total duration of each one of the four different sessions (Helicopters: one vs. two x Language: Greek vs. English) and different conditions (cruising vs. landing). The PPG signal was processed using the HeartPy Python package (van Gent et al., 2018a, 2018b), especially developed for dealing with noisy data collected by commercial wearable ring devices such as the EmotiBit device. We applied the cleaning method provided by the toolkit, using cutoff frequencies of 0.8 Hz and 2.5 Hz, allowing for heart rates from 48 bpm to 150 bpm. The signal was then up sampled to 1000 Hz using the provided method, and peak position was cleaned with an outlier rejection. Time-domain measures were computed and the analysis focused on the Heart Rate Variability (HRV) metric of cardiac activity (i.e., the standard deviation of intervals between adjacent heartbeats, SDNN) measured for the total duration of the session. Fig. 9 shows the mean HRV measured at the two distinct flight phases, for all session types. No significant effects of workload on HRV were found [$F(1,11) = 1.925$, $p = .193$, $\eta^2 = 0.033$].

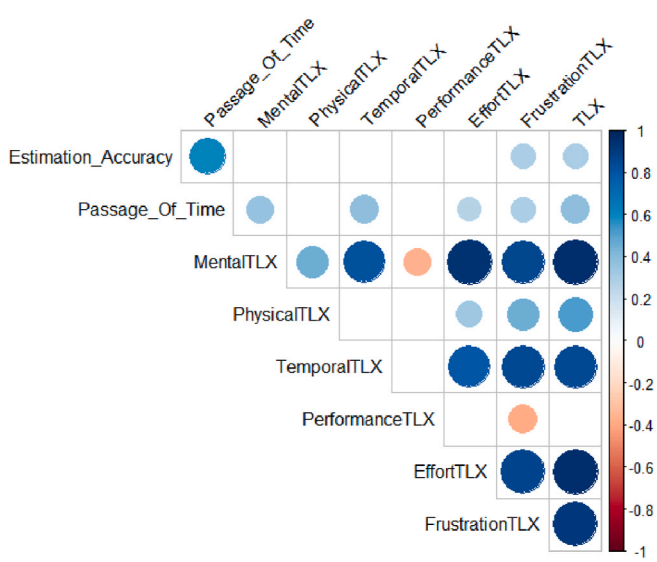

Fig. 7. Correlations between Timing Behavior and Subjective Assessments of Workload. *Note.* Correlations between Estimation Accuracy, Passage of Time Judgments, and the NASA TLX individual and combined score. Only statistically significant correlations are shown. Positive correlations are displayed in blue and negative correlations in red color. The color intensity and the size of the circle are proportional to the correlation coefficients. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

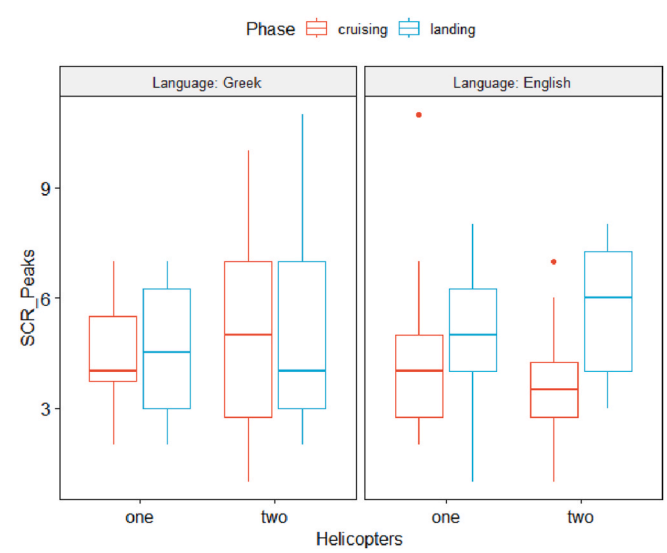

Fig. 8. Mean Skin Conductance Response Peaks at the different sessions and session conditions. *Note.* Mean number of Skin Conductance Response Peaks (SCR_Peaks) for the different flight phase condition (cruising vs. landing), number of helicopters (one vs. two) and language of communication (Greek vs. English). The error bars represent the standard error of the mean.

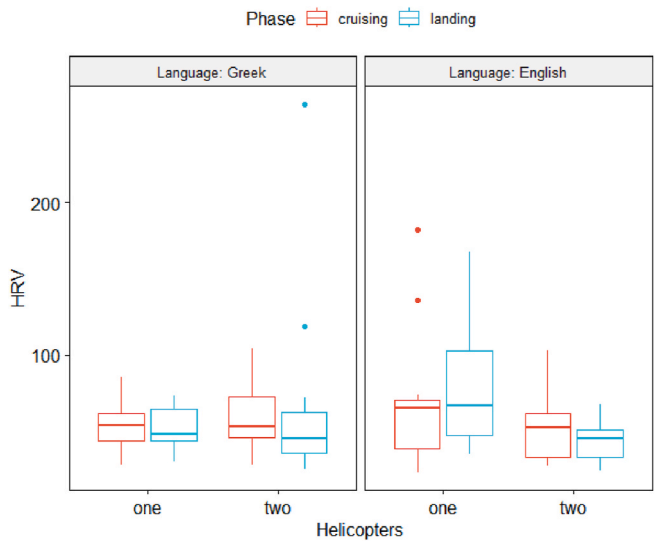

Fig. 9. Mean Heart Rate Variability (HRV) for the different session conditions. *Note.* Mean heart rate variability measured for the different flight phases and session types. The error bars represent the standard error of the mean.

4. Discussion

In the present study, we utilized a dual-task paradigm set up in a real working environment, where we manipulated the load of the non-timing, flight control task (number of helicopters, language of communication, flight phase) to examine the potential effect on the performance (production accuracy, estimation accuracy, passage of time judgment) in the timing task. Additionally, we evaluated the relation between the subjective assessments of workload (NASA TLX) and physiological activity, and the timing behavior. Our results showed that increased workload in the flight task led to overproduction of intervals in the concurrent production task, overestimation of time in the delayed estimation task, and faster passing of time judgments. Furthermore, subjective workload scores were positively correlated with the estimation accuracy and the passage of time judgments, while skin conductance responses were positively correlated with production accuracy.

The current study assessed whether workload manipulations would affect the timing behavior of air traffic controllers in a real aviation environment. Previous studies performed in strictly controlled setups have shown that changes in workload lead to changes in duration judgments (for a review see Block et al., 2010; Matthews and Meck, 2016), with similar findings being reported for semi-controlled settings like simulations setups (Baldauf et al., 2009; Casali and Wierwille, 1983; Hart et al., 1978; Wierwille et al., 1985; Wierwille and Connor, 1983; Zakay and Shub, 1998). Research on time perception has acknowledged that laboratory studies differ significantly from the real world (Matthews and Meck, 2014). The use of simple, homogenous, or static stimuli and the artificial nature of the tasks employed to measure timing behavior bears little resemblance to everyday situations, or highly demanding working environments, where decision making is an important process. Results obtained by timing repeatedly presenting, well-defined in terms of start and end stimuli (van Rijn, 2018) cannot be straightforwardly extended in complex setups, where multiple, dynamically changing events may be timed simultaneously (Matthews and Meck, 2014). Our results are in line with the basic experimental findings of underestimation of time in prospective paradigms, and overestimation of time retrospectively, under high workload conditions, confirming the hypothesis that the theoretical basic research can be generalized to more naturalistic, real-world environments.

Furthermore, when examining the subjective time estimations of the controllers, we found that these significantly differed from the objective time when workload increased. The operators underestimated time, while they were performing the assigned tasks with higher load. That is, they experienced a "shortening" of time as the task demands increased. Retrospectively, those same sessions of increased load were overestimated, with the controllers experiencing a "lengthening" of time. The implications of such an effect of workload on the subjective experience of time are important when considering dynamic environments like air traffic control, where operators continuously adapt their strategies in order to cope with the task demands. Architectural models of workload have identified time as an important factor for managing workload, mainly in the form of time stress or time pressure. Sperandio (1971) sees temporal stress as one of the variables that can change the operative methods used by operators, which, in turn, can modify the workload imposed on the operator. In the information processing model for predicting workload and performance (Hendy et al., 1997), it is suggested that time pressure, seen as the ratio of the time required to make a decision to the time available to solve a problem, causes the operator to adapt their strategy, thus translating to the experienced workload. Acknowledging the subjective nature of time, Hollnagel (2002) proposes that the level and nature of control over the work progress depends on the operator's subjective estimation of the time needed to evaluate events, select, and perform an action, relative to the subjective assessment of the available time. Though a strong correlation between objective and subjective available time is assumed (Hollnagel, 1998), the results of our experiment show that this relation depends on the workload conditions, suggesting a bidirectional relation between perceived workload and time pressure. Also interesting is the fact that even when asked to estimate after-the-fact the duration of flight sessions, which experienced controllers have performed multiple times, and even though the sessions consisted of relatively simple sequence of events, air traffic controllers overestimated time by almost 50% in the case of two helicopters, but underestimated time by 16% in the case of one helicopter. This suggests that workload affects the estimated time required to perform a sequence of events, when these estimations happen under load pressure. Similarly, underestimations found during the performance of the high-load flight control sessions suggest that the operators feel that they have less time available to act, which increases the subjective time pressure.

Though using the performance in a secondary task as a workload index is considered a valid method, attempts to apply it in a real-world scenario are lacking, even though it can be used for immediate detection and alleviation of possible decline in performance in the primary task. This is mainly due to safety concerns, given that the secondary task may interfere with the primary task, or not be sensitive enough to detect changes in the workload (Hart and Wickens, 1990; Meshkati and Loewenthal, 1988; Paas et al., 2003). The secondary task in our experiment was a time production task. It has been found that the use of a timing task as secondary exhibits an asymmetrical interference effect in most cases (for a review see Brown, 1997, 2006), thus, being affected by the primary task but not affecting it. Specific experiments examining the effect of the timing task in piloting performance in flight simulators (Casali and Wierwille, 1983; Wierwille and Connor, 1983; Zakay and Shub, 1998), verified the lack of bidirectional interference between the timing and non-timing task. In a study that found that timing task did affect the performance in piloting (Wierwille et al., 1985), the primary task involved performing mental arithmetic (i.e., solve trigonometric problems) of varied degrees of difficulty, a task that is known to be affected by the timing task. Additionally, it is suggested that if the secondary task is incorporated in the primary task in a normal way (Cain, 2007), this will minimize the unwanted safety hazard issues. If the secondary task is part of the primary task routine it can be performed with minimal intrusiveness. In our experiment, we chose to use a common hardware equipment both for the flight and the temporal production task, the communications digital recorder, which the air traffic controllers are trained to utilize. Apart from being unobtrusive, it has been found that duration judgments and especially prospective time production tasks are rather sensitive to detecting changes in workload.

Block et al. (2010) found through a meta-analysis that prospective temporal productions were affected the most by changes in workload, thus being the more sensitive to them. Overall, the time production task used in our experiment seems to be fulfilling all requirements to be incorporated in a real aviation environment.

In our experiment, the cognitive load imposed on the controller was manipulated both across and within sessions. Different sessions were created by changing the number of helicopters (one vs. two) and the language of communication between the air traffic controller and the pilot (native: Greek, non-native: English), thus, creating four session types. Within each session two different flight phases were identified (cruising vs. landing phase). For the production task, all three factors are considered as possible modulators, while for the estimation task and passage of time judgment only the first two are relevant since each session encompassed both flying phases (cruising and landing). The results obtained showed that all metrics (production accuracy, estimation accuracy, and passage of time rankings) were affected by the number of helicopters. The number of aircrafts under control has been shown to affect the workload (Edwards et al., 2017), suggesting that differences in the time behavior of the controllers across and within sessions were due to the manipulation of workload through the number of helicopters, as hypothesized. Contrary to our hypothesis, changing the language of communication did not have any effect on the duration and passage of time judgments. This finding suggests that using a non-native language did not modulate the workload sufficiently enough to affect the timing behavior of the controllers. This may be attributed to the nature of the language used in air traffic control (Hopkin, 1982), which mainly consists of phrases with specific terms rather than full sentences, the situation (typical flight vs. abnormal events) at which the communication takes part (Prinzo et al., 2010), and the degree of knowledge of the non-native language. The results suggest that short messages with a strict terminology, used in routine flight sessions, spoken by expert, non-native language speakers did not increase the workload sufficiently enough to cause changes in timing behavior. Production accuracy was also affected by the stage at which it was made, that is, whether it was concurrent with a cruising or landing helicopter. Varying the flight phase has been shown to increase workload (Edwards et al., 2017), suggesting that our hypothesis that differences in the production times within a session could be attributed to differences in workload between the two flight phases, is confirmed.

Subjective assessments of task load as depicted in the NASA TLX scores corresponded to the expected workload levels, and they were correlated with the timing behavior. This finding, along with the finding that electrodermal activity measured by the number of skin conductance responses also correlated with the production times suggests that timing performance, subjective assessments, and physiological measures can act as workload measurement techniques (Cain, 2007). In our experiment, no significant effects of workload manipulations on either cardiac (heart rate variability) or electrodermal activity (skin conductance responses) were observed. Though physiological measurements have been used as a load-measuring technique, alongside task performance and subjective rankings, recent studies suggest that there are many factors that can decrease their sensitivity and validity in measuring workload (for a review see (Ayres et al., 2021). Heart rate variability is rather task-specific, showing low sensitivity in simulation-like tasks (Ayres et al., 2021), or not well-controlled setups (Paas et al., 1994). Furthermore, heart rate variability can detect large differences of workload but is less prone to more subtle ones. This may explain the lack of results in our experiment. Across the four types of sessions (one vs. two helicopters x Greek vs. English language), workload manipulation may not have been strong enough to be detected by heart rate variability. Timing behavior was sensitive to the Helicopters factor, and subjective (NASA TLX) rankings detected both the Helicopter and the Language factor. Similarly, electrodermal activity, though affected by workload (Mehler et al., 2012; Nourbakhsh et al., 2017; Setz et al., 2010), is also closely related to stress, making it a confounding factor. In our experiment, skin conductance responses were used to detect changes in the workload between the two different flight phases and across the different session types, which may more probably contain the sudden changes that skin conductance response metrics can detect (Charles and Nixon, 2019). The fact that no effect of workload on skin conductance responses was found, but a correlation to the production times was revealed suggests that there is an effect of load on skin conductance, but our experimental data were not able to reveal it.

In conclusion, the present study showed that timing performance is modulated by changes in workload in a real aviation environment, making timing behavior a valid and sensitive index in detecting variations in the cognitive demands of a task, along with the subjective assessments of difficulty, and the physiological measurements. Air traffic controllers exhibited a differential temporal percept, depending on the number of helicopters they controlled and the flight phase of the helicopters, but were unaffected by the language of communication. Future work could address the limitations of the current study, by identifying more discrete flight phases, apart from cruising or landing, such as take-off, and additional or urgent communication situations. Designing more complex scenarios, that include sequential events and sessions of larger durations and which are commonly encountered by air traffic controllers, and monitoring timing awareness at different points, could reveal the effect of dynamically structured stimuli on time performance. Furthermore, such an approach could potentially reveal the effect of language, especially when a non-expected interaction with the pilot is required, but also further investigate the effect of states with alternating workload. This would also provide us with more data (behavioral and physiological) that could be used to examine the convergent validity between the different workload measures. Especially in the case of physiological data, longer sessions, with sequential manipulations of workload, and the introduction of a baseline period could help us reveal how these are affected by workload changes.

Open practices

Analysis code and research materials are available at https://osf.io/4vtgc/. Data were analyzed using R, version 4.0.0 (R Core Team, 2020). The study was not pre-registered.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964464. The authors would like to thank V. Karadima for her help on the physiological data capture and Dr. C. Parpoula for her advice on the statistical analysis of the data. We would like to thank the Reviewers for their time and effort, and we appreciate all the insightful comments and suggestions for improving the quality of the manuscript. Part of this work was presented as AP's master's thesis.

References

Antonenko, P., Paas, F., Grabner, R., van Gog, T., 2010. Using electroencephalography to measure cognitive load. Educ. Psychol. Rev. 22 (4), 425–438. https://doi.org/10.1007/s10648-010-9130-y.

Ayres, P., Lee, J.Y., Paas, F., van Merriënboer, J.J.G., 2021. The validity of physiological measures to identify differences in intrinsic cognitive load. Front. Psychol. 12. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.702538.

Baldauf, D., Burgard, E., Wittmann, M., 2009. Time perception as a workload measure in simulated car driving. Appl. Ergon. 40 (5), 929–935. https://doi.org/10.1016/j.apergo.2009.01.004.

Block, R.A., Hancock, P.A., Zakay, D., 2010. How cognitive load affects duration judgments: a meta-analytic review. Acta Psychol. 134 (3), 330–343. https://doi.org/10.1016/j.actpsy.2010.03.006.

Boltz, M., 2005. Duration judgments of naturalistic events in the auditory and visual modalities. Percept. Psychophys. 67 (8), 1362–1375. https://doi.org/10.3758/BF03193641.

Brown, S.W., 1995. Time, change, and motion: the effects of stimulus movement on temporal perception. Percept. Psychophys. 57 (1), 105–116. https://doi.org/10.3758/BF03211853.

Brown, S.W., 1997. Attentional resources in timing: interference effects in concurrent temporal and nontemporal working memory tasks. Percept. Psychophys. 59 (7), 1118–1140. https://doi.org/10.3758/BF03205526.

Brown, S.W., 2006. Timing and executive function: bidirectional interference between concurrent temporal production and randomization tasks. Mem. Cognit. 34 (7), 1464–1471. https://doi.org/10.3758/BF03195911.

Cain, B., 2007. A Review of the Mental Workload Literature. https://apps.dtic.mil/sti/citations/ADA474193.

Cardosi, K.M., Murphy, E., 1995. *Human Factors in the Design and Evaluation of Air Traffic Control Systems* (DOT-VNTSC-FAA-95-3). https://rosap.ntl.bts.gov/view/dot/8708.

Casali, J.G., Wierwille, W.W., 1983. A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. Hum. Factors 25 (6). https://doi.org/10.1177/001872088302500602.

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.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S.Z., Khawaji, A., Conway, D., 2016. Robust Multimodal Cognitive Load Measurement. Springer International Publishing. https://doi.org/10.1007/978-3-319-31700-7.

Edwards, T., Gabets, C., Mercer, J., Bienert, N., 2017. Task demand variation in air traffic control: Implications for workload, fatigue, and performance 484, 91–102. https://doi.org/10.1007/978-3-319-41682-3_8.

Estival, D., Farris, C., Molesworth, B., 2016. Aviation English: A Lingua Franca for Pilots and Air Traffic Controllers. Routledge. https://doi.org/10.4324/9781315661179.

Farris, C., Trofimovich, P., Segalowitz, N., Gatbonton, E., 2008. Air traffic communication in a second language: implications of cognitive factors for training and assessment. Tesol Q. 42 (3), 397–410. https://doi.org/10.1002/j.1545-7249.2008.tb00138.x.

Gjoreski, M., Luštrek, M., Pejović, V., 2018. My watch says I'm busy: inferring cognitive load with low-cost wearables. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 1234–1240. https://doi.org/10.1145/3267305.3274113.

Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J.M., Van den Bergh, O., 2016. Respiratory changes in response to cognitive load: a systematic review. Neural Plast. 2016 https://doi.org/10.1155/2016/8146809.

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.

Hart, S.G., Mcpherson, D., Loomis, L.L., 1978. Time Estimation as a Secondary Task to Measure Workload: Summary of Research. https://ntrs.nasa.gov/citations/19790007463.

Hart, S.G., Wickens, C.D., 1990. Workload assessment and prediction. In: Booher, H.R. (Ed.), Manprint. Springer Netherlands, pp. 257–296. https://doi.org/10.1007/978-94-009-0437-8_9.

Hasegawa, M., Carpenter, P.A., Just, M.A., 2002. An fMRI study of bilingual sentence comprehension and workload. Neuroimage 15 (3), 647–660. https://doi.org/10.1006/nimg.2001.1001.

Hendy, K.C., Liao, J., Milgram, P., 1997. Combining time and intensity effects in assessing operator information-processing load. Hum. Factors: The Journal of the Human Factors and Ergonomics Society 39 (1), 30–47. https://doi.org/10.1518/001872097778940597.

Hilburn, B., 2004. Cognitive Complexity in Air Traffic Control: A Literature Review.

Hollnagel, E., 1998. Context, Cognition, and Control. Co-Operation in Process Management-Cognition and Information Technology.

Hollnagel, E., 2002. Time and time again. Theor. Issues Ergon. Sci. 3 (2), 143–158. https://doi.org/10.1080/14639220210124111.

Hopkin, V.D., 1982. Human Factors in Air Traffic Control, first ed. CRC Press. https://doi.org/10.1201/9780203751718.

Hurst, M.W., Rose, R.M., 1978. Objective job difficulty, behavioural response, and sector characteristics in air route traffic control centres. Ergonomics 21 (9), 697–708. https://doi.org/10.1080/00140137808931772.

Jaiswal, D., Chatterjee, D., Gavas, R., Ramakrishnan, R.K., Pal, A., 2021. Effective assessment of cognitive load in real-world scenarios using wrist-worn sensor data. In: Proceedings of the Workshop on Body-Centric Computing Systems, pp. 7–12. https://doi.org/10.1145/3469260.3469666.

Joint Report: KLM-PAA, 1978. Subsecretaria de Aviacion Civil (Aircraft Accident Report: PAN American World Airways Boeing 747, N 737 PA, KLM Royal Dutch Airlines Boeing 747, PH-BUF, Tenerife, Canary Islands). http://www.project-tenerife.com/engels/PDF/Tenerife.pdf.

Kontogiannis, T., Malakis, S., 2017. Cognitive engineering and safety organization in air traffic management. In: Cognitive Engineering and Safety Organization in Air Traffic Management. https://doi.org/10.1201/b22178.

Lin, Y., 2021. Spoken instruction understanding in air traffic control: challenge, technique, and application. Aerospace 8 (3). https://doi.org/10.3390/aerospace8030065. Article 3.

Liu, Y., Wickens, C.D., 1994. Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. Ergonomics 37 (11), 1843–1854. https://doi.org/10.1080/00140139408964953.

Longo, L., Wickens, C.D., Hancock, G., Hancock, P.A., 2022. Human mental workload: a survey and a novel inclusive definition. Front. Psychol. 13 https://doi.org/10.3389/fpsyg.2022.883321.

Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., Chen, S.H.A., 2021. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav. Res. Methods 53 (4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y.

Matthews, W.J., Meck, W.H., 2014. Time perception: the bad news and the good. Wiley Interdisciplinary Reviews Cognitive Science 5, 429–446. https://doi.org/10.1002/wcs.1298.

Matthews, W.J., Meck, W.H., 2016. Temporal cognition: connecting subjective time to perception, attention, and memory. Psychol. Bull. 142 (8), 865–907. https://doi.org/10.1037/bul0000045.

Mehler, B., Reimer, B., Coughlin, J.F., 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. Hum. Factors 54 (3), 396–412. https://doi.org/10.1177/0018720812442086.

Meshkati, N., Loewenthal, A., 1988. An eclectic and critical review of four primary mental workload assessment methods: a guide for developing a comprehensive model. In: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology, vol. 52. North-Holland, pp. 251–267. https://doi.org/10.1016/S0166-4115(08)62391-2.

Mogford, R.H., Murphy, E.D., Roske-Hofstrand, R.J., Yastrop, G., Guttman, J.A., 1994. Application of Research Techniques for Documenting Cognitive Processes in Air Traffic Control: Sector Complexity and Decision Making. https://apps.dtic.mil/sti/citations/ADA282336.

Nourbakhsh, N., Chen, F., Wang, Y., Calvo, R., 2017. Detecting users' cognitive load by galvanic skin response with affective interference. ACM Transactions on Interactive Intelligent Systems 7, 1–20. https://doi.org/10.1145/2960413.

O'Donnell, R.D., Eggemeier, F.T., 1986. Workload assessment methodology. In: Handbook of Perception and Human Performance. Cognitive Processes and Performance, vol. 2. John Wiley & Sons, pp. 1–49.

Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.M., 2003. Cognitive load measurement as a means to advance cognitive load theory. Educ. Psychol. 38 (1), 63–71. https://doi.org/10.1207/S15326985EP3801_8.

Paas, F., Van Merrienboer, J.J.G., Adam, J., 1994. Measurement of cognitive load in instructional research. Percept. Mot. Skills 79, 419–430. https://doi.org/10.2466/pms.1994.79.1.419.

Pape, A.M., Wiegmann, D.A., Shappell, S., 2001. Air Traffic Control (ATC) Related Accidents and Incidents: A Human Factors Analysis.

Prinzo, O.V., Campbell, A., Hendrix, A.M., Hendrix, R., 2010. U.S. Airline Transport Pilot International Flight Language Experiences. American Psychological Association. *Report 4: Non-native English-speaking controllers communicating with native English-speaking pilots*. https://doi.org/10.1037/e733882011-001.

Riemer, M., Wolbers, T., van Rijn, H., 2021. Age-related changes in time perception: the impact of naturalistic environments and retrospective judgements on timing performance. Q. J. Exp. Psychol. 74 (11), 2002–2012. https://doi.org/10.1177/17470218211023362, 2006.

Roitsch, P.A., Babcock, G.L., Edmunds, W.W., 1977. *Human factors report on the Tenerife accident* (aircraft accident report: PAN American world airways boeing 747, N 737 PA, KLM royal Dutch airlines boeing 747, PH-BUF, tenerife, canary islands). In: Air Line Pilots Association, Engineering and Air Safety. http://www.project-tenerife.com/engels/PDF/alpa.pdf.

Romine, W.L., Schroeder, N.L., Graft, J., Yang, F., Sadeghi, R., Zabihimayvan, M., Kadariya, D., Banerjee, T., 2020. Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: development of a cognitive load tracker for both personal and classroom use. Sensors 20 (17). https://doi.org/10.3390/s20174833. Article 17.

Rosner, B., 1983. Percentage points for a generalized ESD many-outlier procedure. Technometrics 25 (2), 165–172. https://doi.org/10.2307/1268549.

Rubio, S., Díaz, E., Martín, J., Puente, J.M., 2004. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Appl. Psychol.: Int. Rev. 53, 61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x.

Schlichting, N., Damsma, A., Aksoy, E.E., Wächter, M., Asfour, T., van Rijn, H., 2018. Temporal context influences the perceived duration of everyday actions: assessing the ecological validity of lab-based timing phenomena. J. Cogn. 2 (1), 1. https://doi.org/10.5334/joc.4.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., Ehlert, U., 2010. Discriminating stress from cognitive load using a wearable EDA device. IEEE Trans. Inf. Technol. Biomed. 14 (2), 410–417. https://doi.org/10.1109/TITB.2009.2036164.

Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F., 2007. Galvanic skin response (GSR) as an index of cognitive load. In: CHI '07 Extended Abstracts on Human Factors in Computing Systems, pp. 2651–2656. https://doi.org/10.1145/1240866.1241057.

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.

Stein, E., 1985. Air Traffic Controller Workload: an Examination of Workload Probe. Federal Aviation Administration. DOT/FAA/CT-TN84/24).

Sucala, M., Scheckner, B., David, D., 2011. Psychological time: interval length judgments and subjective passage of time judgments. Curr. Psychol. Lett. Behav. Brain & Cogn. 26 (2) https://doi.org/10.4000/cpl.4998.

Tachmatzidou, O., Vatakis, A., 2023. Attention and schema violations of real world scenes differentially modulate time perception. Sci. Rep. 13 (1) https://doi.org/10.1038/s41598-023-37030-2.

Tiewtrakul, T., Fletcher, S., 2010. The challenge of regional accents for aviation English language proficiency standards: a study of difficulties in understanding in air traffic control-pilot communications. Ergonomics 53, 229–239. https://doi.org/10.1080/00140130903470033.

Tobin, S., Bisson, N., Grondin, S., 2010. An ecological approach to prospective and retrospective timing of long durations: a Study involving gamers. PLoS One 5 (2), e9271. https://doi.org/10.1371/journal.pone.0009271.

Tsang, P.S., Vidulich, M.A., 2006. Mental workload and situation awareness. In: Handbook of Human Factors and Ergonomics, third ed. John Wiley & Sons, Inc, pp. 243–268. https://doi.org/10.1002/0470048204.ch9.

Tse, P., 2004. Attention and the subjective expansion of time. Percept. Psychophys. 66 (7), 1171–1189. https://doi.org/10.3758/BF03196844.

van Gent, P., Farah, H., Nes, N., Arem, B., 2018a. Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the Fast Lane project. https://doi.org/10.13140/RG.2.2.24895.56485.

van Gent, P., Farah, H., Nes, N., Arem, B., 2018b. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In: Proceedings of the 6th HUMANIST Conference, pp. 173–178.

Van Orden, K.F., Limbert, W., Makeig, S., Jung, T.P., 2001. Eye activity correlates of workload during a visuospatial memory task. Hum. Factors 43 (1), 111–121. https://doi.org/10.1518/001872001775992570.

Van Rijn, H., 2018. Towards ecologically valid interval timing. Trends Cognit. Sci. 22 (10), 850–852. https://doi.org/10.1016/j.tics.2018.07.008.

Van Roon, A.M., Mulder, L.J.M., Althaus, M., Mulder, G., 2004. Introducing a baroreflex model for studying cardiovascular effects of mental workload. Psychophysiology 41 (6), 961–981. https://doi.org/10.1111/j.1469-8986.2004.00251.x.

Watt, J.D., 1991. Effect of boredom proneness on time perception. Psychol. Rep. 69 (1), 323–327. https://doi.org/10.2466/pr0.1991.69.1.323.

Wearden, J.H., 2005. The wrong tree: time perception and time experience in the elderly. In: Duncan, J., Phillips, L., McLeod, P. (Eds.), Measuring the Mind: Speed, Control, and Age. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198566427.003.0006.

Wearden, J.H., Edwards, H., Fakhri, M., Percival, A., 1998. Why "sounds are judged longer than lights": application of a model of the internal clock in humans. Q. J. Exp. Psychol. Sect. B Comp. Physiol. Psychol. 51 (2), 97–120. https://doi.org/10.1080/71393267. Scopus.

Wierwille, W.W., Connor, S.A., 1983. Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. Hum. Factors 25 (1), 1–16. https://doi.org/10.1177/001872088302500101.

Wierwille, W.W., Eggemeier, F.T., 1993. Recommendations for mental workload measurement in a test and evaluation environment. Hum. Factors 35 (2), 263–281. https://doi.org/10.1177/001872089303500205.

Wierwille, W.W., Rahimi, M., Casali, J.G., 1985. Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. Hum. Factors 27 (5), 489–502. https://doi.org/10.1177/001872088502700501.

Wilcox, R.R., Schönbrodt, F.D., 2014. The WRS Package for Robust Statistics in R (0.24) [Computer software]. https://r-forge.r-project.org/projects/wrs/.

Wu, Y., Edwards, J., Cooney, O., Bleakley, A., Doyle, P.R., Clark, L., Rough, D., Cowan, B. R., 2020. Mental workload and language production in non-native speaker IPA interaction. In: Proceedings of the 2nd Conference on Conversational User Interfaces, pp. 1–8. https://doi.org/10.1145/3405755.3406118.

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.

Zakay, D., 1993. Time estimation methods—do they influence prospective duration estimates? Perception 22 (1), 91–101. https://doi.org/10.1068/p220091.

Zakay, D., Fallach, E., 1984. Immediate and remote time estimation—a comparison. Acta Psychol. 57 (1), 69–81. https://doi.org/10.1016/0001-6918(84)90054-4.

Zakay, D., Shub, J., 1998. Concurrent duration production as a workload measure. Ergonomics 41 (8), 1115–1128. https://doi.org/10.1080/001401398186423.