Assessment of Ocular and Physiological Metrics to Discriminate Flight Phases in Real Light Aircraft

Sébastien Scannella, **Vsevolod Peysakhovich**, **Florian Ehrig**, Université de Toulouse, France, **Evelyne Lepron**, Emosciences, Toulouse, France, and **Frédéric Dehais**, Université de Toulouse, France

Objective: The purpose of the present study was to find psychophysiological proxies that are straightforward to use and could be implemented in actual flight conditions to accurately discriminate pilots' workload levels.

Background: Piloting an aircraft is a complex activity where cognitive limitations may jeopardize flight safety. There is a need to implement solutions to monitor pilots' workload level to improve flight safety. There has been recent interest in combining psychophysiological measurements. Most of these studies were conducted in flight simulators at the group level, limiting the interpretation of the results.

Methods: We conducted an experiment with 11 pilots performing two standard traffic patterns in a light aircraft. Five metrics were derived from their ocular and cardiac activities and were evaluated through three flight phases: takeoff, downwind, and landing.

Results: Statistical analyses showed that the saccadic rate was the most efficient metric to distinguish between the three flight phases. In addition, a classifier trained on the ocular data collected from the first run predicted the flight phase within a second run with an accuracy of 75%. No gain in the classifier accuracy has been found by combining cardiac and ocular metrics.

Conclusions: Ocular-based metrics may be more suitable than cardiac ones to provide relevant information on pilots' flying activity in operational settings.

Applications: Electrocardiographic and eye-tracking devices could be implemented in future cockpits as additional flight data for accident analysis, an objective pilot's state evaluation for training, and proxies for human-machine interactions to improve flight safety.

Keywords: workload, aircraft pilots, eye-tracker, ECG, classification.

Address correspondence to Sébastien Scannella, ISAE-SUPAERO, Université de Toulouse, 10 avenue Edouard Belin, Toulouse, 31056, France; e-mail: s.scannella@isae-supaero.fr.

INTRODUCTION

Operating an aircraft is a challenging task. Pilots have to monitor numerous flight deck gauges, communicate with air traffic controllers, and make decisions to adapt to external contingencies to ensure flight safety. For instance, during the takeoff and landing phases, pilots have to monitor the speed, attitude, and altitude of the aircraft; adjust thrust power; and control flaps' position, which may induce high workload levels (Dehais, Behrend, Peysakhovich, Causse, & Wickens, 2017). This excessive task demand can lead to attentional impairments (Dehais, Causse, Vachon, & Tremblay, 2012; Dehais et al., 2014; Dehais, Tessier, Christophe, & Reuzeau, 2010; Thomas & Wickens, 2004; Wickens & Alexander, 2009) and consequent persistence in erroneous decisions (Dehais, Tessier, et al., 2010; Reynal, Rister, Scannella, Wickens, & Dehais, 2017). During a cruising phase and under normal circumstances, however, the demand involves considerably less effort. Pilots have mainly to handle radio communications and correct for possible deviations that represent a relative constant low task demand. Hence, it may lead to low vigilance, distraction, mind wandering, and failure to adequately monitor the flight deck (Casner & Schooler, 2014; Durantin, Dehais, & Delorme, 2015; Gouraud, Delorme, & Berberian, 2017). It seems, therefore, that these two extreme levels of workload during a flight may be associated with a decrease in performance that may jeopardize flight safety.

Workload Assessment in Simulated Flights

Decades of research in flight simulators have shown that electrocadiography (ECG) is a reliable approach to derive changes in the activity of the autonomous system (ANS) as an indicator of mental workload variation (Blix, Stromme, & Ursin, 1974; Lee & Liu, 2003; Opmeer & Krol, 1973). Among ECG-derived metrics, one of the most commonly used is the heart rhythm (HR) (Dahlstrom, Nahlinder, Wilson, & Svensson, 2011; Hankins & Wilson, 1998; Jorna, 1993; Wilson, 2002). Due to the relative ease of access when recording ECG data, researchers have also explored the sensitivity of the heart rate variability (HRV) to workload (see Togo & Takahashi, 2009) and proposed it as a valuable metric to identify changes in mental activity in the absence of any overall change in rate (Jorna, 1993; Roscoe, 1992). On this basis, some studies have shown that HRV could be a good mental workload indicator (Durantin, Gagnon, Tremblay, & Dehais, 2014; Sauvet et al., 2009; Veltman & Gaillard, 1993), whereas other studies found contradictory results (Opmeer & Krol, 1973; Roscoe, 1992; Wilson, 2002). In an attempt to reach a consensus, most of the current studies continue to evaluate both HR and HRV.

Another promising approach to derive mental workload is to consider the use of eye tracking (Duchowski, 2007). One great interest of this technique is to provide physiological measures such as blink rate (Hughes & Cole, 1998) and pupil diameter (Causse, Peysakhovich, & Fabre, 2016) as well as behavioral metrics such as gaze velocity (Di Stasi et al., 2010), fixation duration (Backs & Walrath, 1992), saccadic rate (Tokuda, Obinata, Palmer, & Chaparro, 2011), and fixation/saccade ratio (Dehais, Peysakhovich, Scannella, Fongue, & Gateau, 2015). For instance, Di Nocera, Camilli, and Terenzi (2007) have demonstrated that the spatial distribution of fixations on the flight deck could well segregate the workload levels related to five flight segments (i.e., departure, climb, cruising, descent, and landing) in expert Instrument Flight Rules (IFR) pilots.

Complementarity of Psychophysiological Metrics

According to this literature, both ECG and ocular-based metrics have been shown to be sensitive to the flight demand. Some authors, however, have suggested that unimodal metrics may provide only limited interpretation of the mental workload. Hankins and Wilson (1998) showed for instance that the heart rate is sensitive to the flight mental demand although limited with regard to determining which event specifically induced HR changes. They proposed complementary measures to draw a more comprehensive picture of the flight mental demands by adding the blink rate, which is more specific to visual demands. Similarly, electroencephalography (EEG) (Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2014) and near infrared spectroscopy (NIRS) (Causse & Matton, 2014; Gateau, Durantin, Lancelot, Scannella, & Dehais, 2015) have also been used to assess the pilot's workload level by providing additional information about the nature of it (e.g., visual load in the occipital lobe).

As an illustration, Wilson (2002) evaluated the heart and blink rates as well as different brain activity frequency bands as a mean of markers of pilots' real flight-related workload across 20 flight phases. Wilson found the heart rate to be sensitive to the workload by discriminating among three groups of segments: high heart rate group (takeoff, touch-and-go, and landing), low heart rate group (mainly IFR segments), and an intermediate heart rate group (all remaining segments). Moreover, Wilson also found that both alpha and delta EEG band activities were sensitive to the workload level associated with these flight phases. Wilson concluded that a concomitant effect over the central (EEG) and peripheral (ECG) nervous systems further emphasized the high cognitive demands of the tasks. Taken all together, these results show that the HR could be a reliable metric to evaluate the workload levels induced by relatively different flight phases and that adding complementary metrics could draw a more complete description of the pilot's workload under realistic settings.

Workload Assessment in Real Flights

To date, compared to simulator studies, less experiments have been conducted in actual flight conditions to evaluate pilots' workload (Dahlstrom et al., 2011; Dehais, Causse, & Pastor, 2008; Hankins & Wilson, 1998; Roscoe, 1992; Veltman, 2002; Wilson, 2002). One main reason is that such experiments have to be approved by the national aviation safety agencies, and they are both time consuming and expensive. Moreover, the flight environment affects the data quality of almost all devices and the subjective levels of mental workload compared to flight simulators. Regarding the data quality, the light variation can affect eye tracking and NIRS measurements, and the electromagnetic field of the engine creates artifacts in the ECG and the EEG signals that are not easy to remove. Finally, these ecological settings may be subjective to unexpected events that could obviously affect the cognitive processes. Hence, researchers have almost never used eye-tracking devices in real flight environments. To the authors' knowledge, only one study was conducted in real flight condition with an eye tracker (Dehais et al., 2008) in which the authors have shown that eye-tracking metrics could segregate well between a degraded flight sequence (engine failure) and a nominal one. Today, the technology has evolved, and hardware are less sensitive to interference. As a consequence, ECG and eye-tracking devices are now good candidates to assess the pilot's workload in real flight environment.

While the aforementioned pilot's workload estimation literature provides some insights on how the flight-related workload level may be measured through physiological devices, it doesn't indicate how each pilot responds to the flight phase demands, limiting the interpretation to the group level only. Yet developing idiosyncratic metrics could be used to design adaptive cockpits and/or safety countermeasures in critical situations.

Workload Classification at the Individual Level

In the attempt to develop subject-specific workload assessment, recent applications using machine learning techniques have been tested with encouraging accuracy using offline (Callan, Durantin, & Terzibas, 2015) and even online processes (Gateau et al., 2015). Some limits for real-life applications can be pointed out, though. First, they used EEG or NIRS devices that are difficult to analyze because of the aforementioned environmental noise. In addition, some of them used advanced techniques that are not easy to apply online (independent component analyses), compared with linear discriminant analyses (LDA) or support vector machine (SVM) for instance. Finally, variability over time of the physiological measures of workload has been pointed out previously (Christensen, Estepp, Wilson, & Russell, 2012), suggesting a limitation in the use of a given classification for a long period. On the basis of these results and the aforementioned work on the pilot's workload classification, it appears of a great importance to evaluate the accuracy of a workload classifier trained with previous data to provide individual workload level assessment. More generally, the validation of such a classifier would pave the way to individual in-flight mental workload assessment on the basis of an idiosyncratic baseline that would need minimum recalibration over the time.

Present Study

We evaluated the possibility of using on-board objective tools to monitor pilot's workload in a real flight environment according to the flight phase. Our main goal was to define a psychophysiological proxy that could be stable enough over time to accurately discriminate between the pilot's workload levels across the flight phases at the individual level.

As exposed previously, among all possible measures, the combination of eye tracking and ECG seems to be one of the most promising in terms of on-board implementation and complementarity. We therefore collected cardiac and ocular data from 11 pilots performing two standard traffic patterns in a Robin DR-400 light aircraft. We focused on three main flight phases, namely, the takeoff leg, the crosswind leg, and the landing leg, across two runs. We derived and combined five different metrics corresponding to the HR, HRV (standard deviation of the R-R interval; STD-RR), fixation duration, visual entropy (randomness in the fixation pattern), and saccadic rate. In addition, pilots fulfilled the NASA Task Load Index questionnaire (NASA-TLX; Hart & Staveland, 1988) at the end of the flight to collect subjective workload evaluations for the three flight segments. We expected the takeoff and landing phases, in which safety margins (i.e., energy, flightpath) are minimum and time pressure is maximum, to induce higher workload than the downwind leg. We thus presume these two flight phases to be rated with higher NASA-TLX scores and induce higher

HR and lower HR variability than the downwind leg. On the opposite, we believed that the downwind leg would induce higher saccadic activity than the takeoff and landing phases. Indeed, the downwind requires shared attention abilities to perform predefined actions (e.g., setting flaps and engine parameters) while supervising the flightpath with regard to different external cues (e.g., villages, landing strip), whereas the landing and takeoff phases induce higher focused attention toward the runway axis and the monitoring of very few critical flight parameters (speed, vertical speed) (Dehais et al., 2008). Finally, we hypothesized that takeoff and landing phases would be difficult to segregate with subjective ratings and ECG metrics as the workload is high in both cases and previous studies have often shown no significant difference between them (Dahlstrom et al., 2011; Di Noccera et al., 2007; Wilson, 2002). However, we anticipated that the ocular metrics could be sensitive to visual activity differences because pilots need to spend more time on the instruments during the takeoff leg (Dehais, Causse, & Pastor, 2010).

We first achieved statistical group analyses for flight-phase discrimination. We then use classification techniques, as initially proposed by the pioneering work of Wilson and Fisher (1991), with the particularity of using data from a first run to classify flight phases in a second run.

MATERIAL AND METHODS

Participants

Eleven healthy volunteers (2 women; mean age = 21.4 ± 3.4 years; mean flight hours = 68 ± 19), all students of the ISAE-Supaero (French Higher Engineering Aerospace School), participated in this study. All reported normal or corrected-to-normal vision and hearing. No participant had a history of cardiac or neurological disease, and as required by aeronautical regulations, no participant was taking psychoactive substances or medication. All the participants gave their written consent after having been informed of the nature of the experiment. They all performed the experiment as part of their pilot training program, and they all did their first solo flight before the experiment. This research complied with the tenets of the Declaration of Heksinki.

Flying Task

The flight mission was to perform two real consecutive standard traffic patterns with a touch-and-go. A touch-and-go consists in landing on the runway without a full stop, applying full throttle once all wheels had touched the ground, followed by a takeoff. All pilots performed two traffic patterns (Run 1 and Run 2). Each traffic pattern, according to the standards of visual flight rules (VFR), consisted of five different flight phases—upwind takeoff leg, crosswind leg, downwind leg, base leg, and final leg, which concluded with a touch-and-go (between the two runs) or a landing (at the end of the second run) (Figure 1).

Due to wind orientation, eight participants underwent the right-hand pattern and three the left-hand pattern. Three flight phases, homogeneous across participants, were selected for analyses (see Figure 1):

Phase 1: Takeoff (period of 60 seconds, starting from power setting or touch-and-go)
Phase 2: Downwind (period of 60 seconds, in the middle of downwind)
Phase 3: Landing (period of 60 seconds, before touch down).

We choose this phase duration as a tradeoff between having enough data points for classification and a reactive enough classifier.

Experimental Procedure

All flights were done on the ISAE-SUPAERO Robin DR400 light aircraft. The experiment was approved by the European Aviation Safety Agency (EASA) permit to fly 2403 2424 2487–EASA 0010011661. The data acquisition system was stored in the baggage compartment and consisted of a computer with connections toward power supply, the intercommunication audio system, the heart rate sensor, and the eye-tracking headset. During each flight, three persons were on board, including the participant, the safety pilot (instructor), and the experimenter in the backseat. After the briefing, the participant boarded the aircraft and was

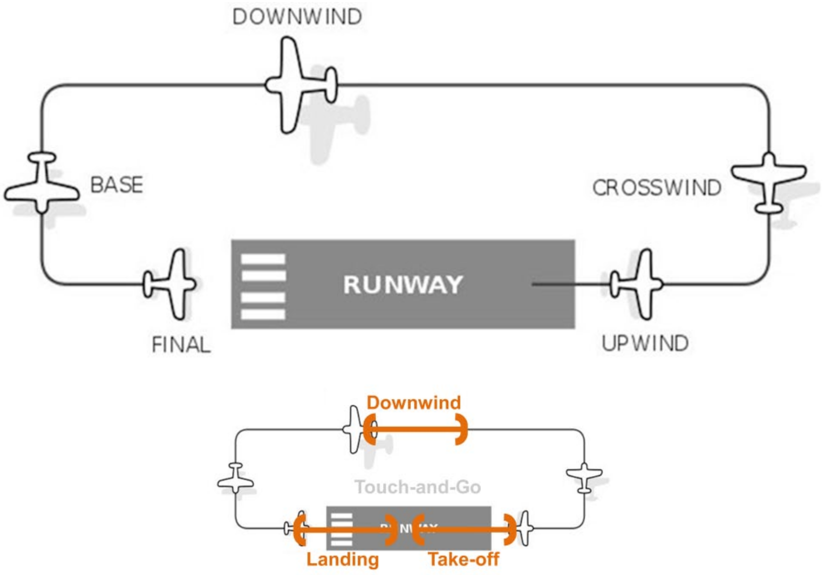

Figure 1. Top: Standard traffic pattern adapted from FAA Aeronautical Information Manual. Bottom: The three flight phases considered for the study are highlighted in orange: takeoff leg, downwind leg, and landing leg, all with a duration of 60 seconds.

connected to the data acquisition system. The installation and calibration of the eye-tracking system lasted 3 minutes and was started after engine start prior to taxi (i.e., running of the plane from the hangar to the runway). Ocular data were recorded from full throttle at takeoff to the moment the aircraft was under control at the final landing. The ECG data recording started at least 30 seconds before and ended 30 seconds after the ocular recordings. Each flight phase marker was noted electronically (via the ECG system) and manually by the experimenter. The experiment was conducted in daylight (late morning, early afternoon) under normal VFR weather conditions. Flight conditions were good and consistent across pilots: clear weather, good visibility, and light wind.

Measurements

Electrocardiography. ECG data were acquired with the ProComp Infinity System (Thought Technology, Montreal, Canada) at a sampling rate of 2048 Hz. Three electrodes were connected to an extender cable and the participant's chest using conductor gel to enhance signal quality (Figure 2). The raw data recorded from the ProComp system were recorded and stored on the laptop computer equipped with Thought Technology recording software.

Oculometry. Eye-tracking data were collected using a head-mounted Pertech eye tracker (Pertech, Mulhouse, France). This device has a nominal accuracy of 0.25° and a maximum sampling rate of 50 Hz. A calibration procedure was performed using five distant points on the instrument flight panel. The data from a field camera and an infrared-sensitive eye camera mounted to the head support were recorded by the computer unit installed in the aircraft baggage compartment. The left pupil barycenter axis and the field camera x-y positions were recorded. To improve data quality and facilitate the viewing for the pilot, a sunscreen was fixed to the head-mounted eye-tracker system (Figure 3). The sunscreen (darkened glasses) was fixed behind the eye-tracking camera (relatively to the participants' eyes) and thus was not obstructing the pupil detection.

Data Analyses

Subjective workload. To compare the sensitivity of our psychophysiological metric to the

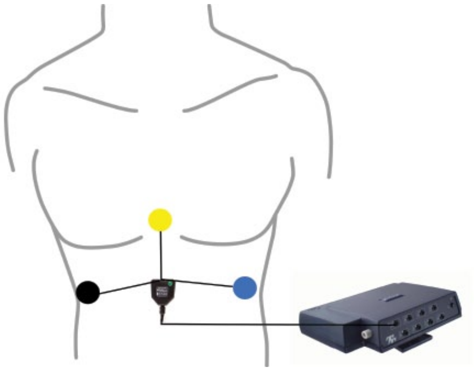

Figure 2. Electrocardiograph ProComp Infinity electrode positions.

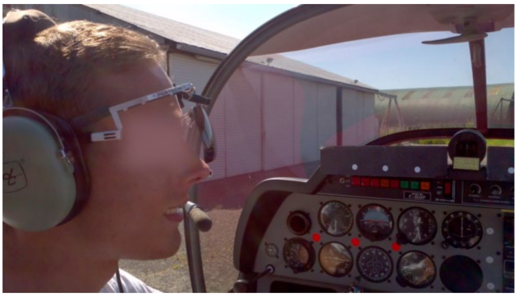

Figure 3. Pilot with eye-tracking setup on board of a DR400 light aircraft.

subjective rating of pilot's workload, all pilots fulfilled a NASA-TLX paper-based questionnaire (Hart & Staveland, 1988) during the flight debriefing. They were asked to self-report their subjective scores for each of the three considered flight phases. Overall rating has been calculated using the bipolar weighting comparing two by two each of the six components.

Electrocardiography. The R-R intervals of the raw ECG signal were detected using the built-in QRS detection algorithm of Kubios HRV software (Tarvainen, Niskanen, Lipponen, Ranta-Aho, & Karjalainen, 2014). All the recordings were manually revised for missed or false positive R-peak detection. We then computed the mean values of HR (in beats per minute) and HRV (assessed as the standard deviation of the inter-beat interval, IBI *SD*) within the 60-second window of each of the three phases of the two runs.

Eye tracking. Similarly, we computed the averaged values of fixation duration, saccadic rate, and visual entropy within the 60-second window of each of the three phases of the two runs. The eye movements were detected using a dispersion-velocity–based algorithm. Gaze displacements with a speed inferior to 30° per second with a dispersion threshold of 1° were considered as fixations. Other samples were considered as saccades. The visual entropy was defined as the Mean Neighbor Index (MNI) according to Di Nocera et al. (2007):

$$MNI = \frac{d(MN)}{d(ran)},$$

where $d(MN)$ represents the dispersion of the coordinates of view focus across time as:

$$d(MN) = \sum \frac{mean(d_{ij})}{N},$$

and $d(ran)$ represents the distance between each focus view (dispersion value) within the time windows, defined as:

$$d(ran) = 0.5\sqrt{\frac{\text{Area of View}}{N}}.$$

Ecological conditions of the present study (light conditions, plane vibrations) impacted the quality of the eye-tracking data. Average data quality per condition was of 88% ± 11% of valid samples (minimum = 75%; maximum = 96%). The missing data were independent of any condition (Friedman ANOVA; $p > .26$) and have been tagged to be excluded from analyses.

Statistical Analyses

All statistical analyses were carried out with Statistica 10 (StatSoft). The normality assumption has been assessed with the Lilliefors test ($p > .2$ in all cases). Multivariate analyses for repeated measures (MANOVAs) were computed over the HR, HRV, and each ocular metrics (fixation length, saccade frequency, and visual entropy) separately. For each analysis, within-subject factors run (No. 1 vs. No. 2) and flight phase (takeoff vs. downwind vs. landing) design was used. NASA-TLX global

TABLE 1: MANOVA Results for ECG and Eye-Tracking Metrics

	Effects	*F*(*df*)	*p* Values	Partial Eta^2	Flight Phase Separation
HR	Run	(1, 10) = 10.93	.008	0.52	a,c
	Flight phase	(2, 20) = 14.13	.009	0.59	
	Run × Flight Phase	(2, 20) = 0.05	.95	0.01	
HRV	Run	(1, 10) = 0.72	.42	0.07	c
	Flight phase	(2, 20) = 5.84	.02	0.37	
	Run × Flight Phase	(2, 20) = 0.20	.77	0.02	
Fixation length	Run	(1, 10) = 0.05	.83	0.01	
	Flight phase	(2, 20) = 3.62	.16	0.27	
	Run × Flight Phase	(2, 20) = 0.07	.47	0.02	
Saccadic rate	Run	(1, 10) = 0.31	.59	0.03	a,b,c
	Flight phase	(2, 20) = 25.10	<.001	0.72	
	Run × Flight Phase	(2, 20) = 0.99	.06	0.09	
MNI	Run	(1, 10) = 0.01	.97	<0.01	a,c
	Flight phase	(2, 20) = 38.10	<.001	0.79	
	Run × Flight Phase	(2, 20) = 3.95	.048	0.28	

Note. ECG = electrocadiography; HR = heart rhythm; HRV = heart rate variability; MNI = Mean Neighbor Index.

[a,b,]Honestly significant difference post hoc comparisons; $p < .01$.

[a]Takeoff versus downwind.

[b]Takeoff versus landing.

[c]Downwind versus landing.

scores were also analyzed in a MANOVA with within-subject factors flight phase (takeoff vs. downwind vs. landing). As pilots reported difficulties to quote differently the two runs, they have been asked to consider averaged runs for each flight phase. The Tukey's honestly significant difference (HSD) test was used for all post hoc comparisons. Significance level was set at $p < .05$ for all analyses.

Linear Discriminant Analyses

Three different linear discriminant analyses (LDAs) were computed separately using R software (version 3.2.3) with the saccadic rate alone, HR alone, or both as features to classify the three flight phases. The LDAs were trained on the first run and tested on the second one using a leave one out cross-validation (one participant out) to evaluate within-subject variability over time. The chance level has been calculated at 48% for $p < .05$ and three-class classification, according to Combrisson and Jerbi's (2015) recommendations using the matlab function *binoinv*.

RESULTS

Statistical results and mean values for cardiac and eye metrics are summarized in Table 1 and Table 2.

Subjective Workload

The multivariate test over the NASA-TLX global scores revealed that the three flight phases induced significant subjective workload differences, $F(2, 20) = 65.3$; $p < .001$ (Figure 4). More precisely, the post hoc tests showed that pilots felt the highest level of workload during the landing phase with a mean global workload index of 57.1 ± 5.6, followed by the takeoff phase (45.4 ± 6.1). Finally, as expected, the pilots experienced the lower level of workload during the downwind phase with a mean index of 33.1 ± 6.8 ($p < .001$ for all comparisons).

TABLE 2: Mean Metric Values for the Three Flight Phases Across the Two Runs

	Flight Phase					
	TO 1	Down 1	Land 1	TO 2	Down 2	Land 2
HR (*bpm*)	103.2 (11.4)	95.4 (10.5)	106.2 (15.9)	100.5 (14.1)	92.5 (11.0)	102.5 (11.9)
HRV (*Std*)	0.045 (0.028)	0.047 (0.016)	0.033 (0.020)	0.041 (0.019)	0.044 (0.016)	0.034 (0.019)
Fixation length (*s*)	1.36 (0.78)	0.71 (0.23)	0.76 (0.28)	1.60 (1.02)	0.69 (0.13)	0.81 (0.16)
Saccadic rate (s^{-1})	1.06 (0.26)	1.56 (0.24)	0.48 (0.35)	1.10 (0.43)	1.50 (0.24)	0.61 (0.42)
MNI (*a.u*)	5.47 (2.22)	8.58 (2.75)	3.54 (1.83)	5.02 (2.17)	8.30 (2.32)	4.32 (2.31)

Note. Values in parentheses are the standard deviations.TO = takeoff; Down = downwind; Land = landing; HR = heart rhythm; HRV = heart rate variability; MNI = Mean Neighbor Index.

Cardiac Activity

The MANOVAs showed that the HR and HRV were both significantly impacted by the flight phase ($p < .01$ and $p < .05$, respectively). The post hoc analyses revealed a lower HR for the downwind compared to the two other phases ($p < .01$ for both comparisons), whereas the landing and the takeoff were not significantly different ($p = .45$). The standard deviation of the inter-beat interval was only different between the downwind and the landing phases ($p < .01$). Finally, the HR was the only cardiac metric that was sensitive to a run effect, with higher values for the first run compared to the second one. This run effect did not interact with the flight phase effect.

Ocular Activity

The MANOVAs carried out over the ocular metrics revealed that the saccadic rate and the visual entropy (MNI) were significantly affected by the flight phase ($p < .001$ for both tests), but no main effect of the run has been found. Overall, the saccade frequency and the visual entropy showed high effect sizes for the flight phase, but the saccade frequency was the only visual metric that allowed a significant separation of all phases, as shown by the post hoc results. The visual entropy (MNI) was sensitive to a Run × Flight Phase interaction. This interaction was due to a significant difference between all flight phases within the first run but no difference between the takeoff and the landing in the second run. Finally, the fixation length did not reveal any difference between the three flight phases.

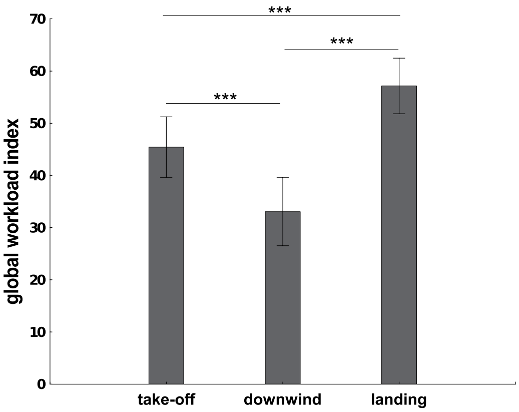

Figure 4. Subjective workload index from the NASA Task Load Index (TLX). The vertical bars represent the standard deviations. ***$p < .001$.

Flight Phase Classification

Among the tested metrics described in the previous section, we selected those that provided the better flight phase separation (i.e., HR for ECG and saccadic rate for eye tracking) to conduct offline LDAs. Data collected during the first run were used to train classifiers dedicated to discriminate the flight phases during the second run. We found that the LDA based on the HR alone provided poor classification accuracy (42%; $p = .17$), whereas the LDA based on the saccadic rate performed far better than the chance level with a global accuracy of 75% ($p < .001$): 64% for takeoff, 75% for downwind, and 86% for landing. Finally, using both ECG and ocular features, we also found 75% of global accuracy ($p < .001$): 61% for takeoff, 80% for downwind, and 84% for landing (see Figure

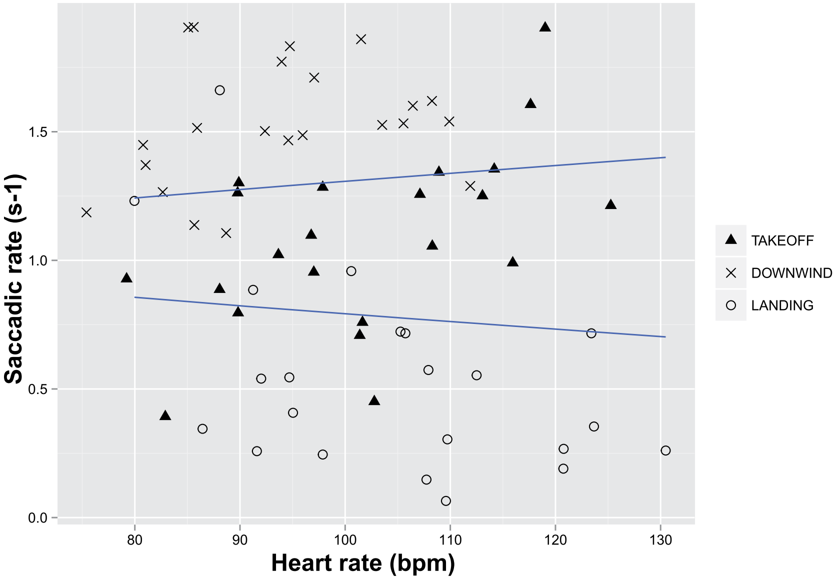

Figure 5. Multivariate linear discriminant analyses (LDA) graphical results for saccadic rate against the heart rate. The LDA was trained on the first run and tested on the second one using a leave one out cross-validation. The blue lines represent the discriminant functions for class separation built with the data of the first run.

5). A MANOVA carried out on the individual results across these three LDA showed that accuracy with the saccadic rate alone was not different than the one obtained with the combined HR and saccadic rate features ($p > .999$). However, the LDA accuracy based on the HR alone was different from the two others ($p < .01$ for both comparisons). At the individual level, the saccadic rate LDA led to a perfect flight phase classification for 6 pilots out of 11, two phases out of three for 4 other pilots, and one phase out of three for 1 remaining pilot.

DISCUSSION

The objective of the present study was to define a psychophysiological proxy that could be sensitive enough over time to accurately discriminate between pilots' workload levels across real flight phases at the individual level. As already shown in previous work (Callan et al., 2015; Dahlstrom et al., 2011; Dehais et al., 2008; Di Nocera et al., 2007; Lee & Liu, 2003; Roscoe, 1992; Wilson, 2002), psychophysiological data for assessment of mental workload can be collected in a real flight environment. To our knowledge, however, it has never been shown that these metrics could provide satisfying mental workload classification accuracy over time.

We choose to focus on two modalities, the ECG and the ocular activity, for their complementarity (Hankins & Wilson, 1998). In addition, both are known to be less sensitive to the surrounding noise within a real flight than other techniques (e.g., EEG and fNIRS) and potentially easy to implement in the flight deck. In our study, pilots were equipped with a head-mounted eye tracker and an ECG sensor. Heart rate and heart rate variability (standard deviation of the inter-beat interval) from the ECG and fixation length, saccadic rate, and visual entropy from the eye tracking were therefore extracted for the takeoff, downwind, and landing flight legs across two standard flying patterns. We found that solely the saccadic rate and the subjective rating allowed a significant separation of the three flight phases at the group level. In addition, a phase classifier based on the saccadic rate of the first run led to a global classification accuracy of 75% within the second run, which performed as well as a multifeature (HR and saccadic rate) classifier.

Group Level

In the present study, we found that all metrics provided a significant flight phase main effect suggesting different mental demands across them, but only the saccadic rate and the NASA-TLX distinguished the three phases from each other. According to the self-reported workload evaluation, pilots felt that the landing was the most demanding phase, followed by the takeoff and downwind. However, when looking at the cardiac activity—which has been often accurately related to pilots' workload level (Dahlstrom et al., 2011; Durantin et al., 2014; Hankins & Wilson, 1998; Jorna, 1993; Sauvet et al., 2009; Veltman & Gaillard, 1993; Wilson, 2002)—neither the HR nor the HRV distinguished the three flight phases. More precisely, the HR was the most accurate of the two and allowed to distinguish between the downwind and the two other phases but not between takeoff and landing. The HRV, meanwhile, solely differentiated between downwind and landing. The limit of these objective workload measurements has already been reported in this kind of environment (Hankins & Wilson, 1998; Wilson, 2002). Wilson (2002) for instance, did not find any difference between takeoff and landing phases either using HR or HRV, probably because these two phases are very close in terms of mental demand. Borghini and colleagues (2014) reported that the HR is also influenced by muscular fatigue, anxiety, and respiration, which are not evaluated in the NASA-TLX. This could arguably account for such difference between subjective and objective measures in the present study. Another interesting result is the nonsignificant interaction between the run and the flight phase that suggests a comparable phase effect between the two runs with global higher HR values for the first run compared with the second one. Thus, despite the occurrence of habituation mechanisms, the relative sensitivity of the HR to the mental workload may be preserved. Overall, the present results of the cardiac analyses provide additional arguments in favor of a better accuracy of the HR compared to the HRV in workload assessment—as already suggested (Dahlstrom et al., 2011; Wilson, 2002). They also point out the need for a complementary measure to more accurately segregate the flight phases and provide a more comprehensive picture of the flight mental demands, especially if one wants to segregate between two tasks with comparable mental demands.

Regarding the eye activity, one important result of this study is the fact that the saccadic rate provided significant post hoc results distinguishing between the three phases at the group level. Thereby, it was the highest for the downwind leg, lowest for the landing leg, and at an intermediate value during takeoff. This can be explained by the fact that pilots need to search for visual references during the downwind leg, whereas they are extremely focused on the runway during landing. The difference between landing and takeoff is more subtle, though. Pilots usually look for instrument values such as the altitude or the speed in both flight phases. Hence, it is likely that the difference comes from longer time spent looking at the runway during the landing, which is associated with lesser saccades.

Based on the HR and saccadic rate results, one could argue that the saccadic rate is more related to the visual demand of the flight phase than the mental workload per se, as it has been already suggested for the blink rate by Hankins and Wilson (1998). Thus, the latter complements the HR in the sense of a description of the task that may have led to a variation in the pilot's workload level. In other words, two comparable objective mental demands can be elicited by two different activities (e.g., the takeoff and landing in our case), but they will not be neither differentiated one from each other nor understood if assessed uniquely with the heart rhythm. Finally, we found mitigated results regarding the eye entropy and the fixation duration. Indeed, the visual entropy—corresponding to how randomly distributed the gaze is in the visual field—allowed separating the three flight phases within the first run, whereas takeoff and landing were no more significantly different in the second run. Hence, only a trend for the global difference (i.e., flight phase principal effect for both runs averaged) between takeoff and landing (p = .057) has been found in the post hoc comparison. Looking closer to the data, whereas almost all pilots exhibited a similar pattern across the two runs, one pilot exhibited a large increase of visual entropy during the landing of the second run. Due to the small number of pilots in this study, it is likely that this value alone could have affected the significance of this result. Nevertheless, looking at Di Nocera and colleagues' (2007) results and the ones presented here, this metric is still promising in terms of online mental workload assessment. The fixation duration, meanwhile, did not lead to any significant effect, making this workload metric not specific enough in the flying domain to segregate well pilots' activity. To summarize, at the group level, the HR and the saccadic rate have been found to be the most sensitive metrics across the cardiac and ocular activities.

Flight Phase Classification

As shown by these results, the limit of unimodal metrics in the workload assessment in such a complex environment has to be taken into account when inferring pilots' activity. If the HR is well known to correlate with pilots' workload levels (Dahlstrom et al., 2011; Hankins & Wilson, 1998; Jorna, 1993; Wilson, 2002), it presents some limited sensitivity that may prevent from distinguishing subtle mental workload changes (e.g., takeoff vs. landing). On the other hand, ocular metrics such as the saccadic rate may discriminate between situations with close mental workload levels but are more related to the visual activity (Hankins & Wilson, 1998). We hence evaluated the accuracy of unimodal flight phase classifiers and the gain in combining these measures. First, the HR-based LDA led to nonsignificant flight phase classification, probably because of a run effect with global higher values in the first run compared with the second one. Indeed, as we wanted to assess the stability of the measure over time, the classifier has been trained with the data of the first run and tested with the data of the second run. This exemplifies previous studies that have suggested variability over the time of psychophysiological data (Christensen et al., 2012). As a consequence, using the cardiac activity to classify pilots' workload online would require at least a baseline recalibration to be more accurate. The LDA with the saccadic rate alone, meanwhile, led to a good accuracy in flight phase classification. Compared to the HR, it seems that segregating the pilot's activity could be better achieved with visual activity than mental demand measurement. Finally, the two-feature LDA—with the saccadic rate and the HR—did not perform better than the classifier with the saccadic rate alone. This result is surprising since we first expected that the complementarity of these measures would help improving the classifier accuracy. Looking closer to the data distribution of the two-feature classifier, it appears that the discriminant functions for class separation are almost parallel to the x-axes, testifying of a null contribution of the HR to the classifier.

Online Workload Estimation

Regarding our prerequisites, the classification of flight phases based on the saccadic rate alone satisfies the required criteria to consider an eye tracker as a useful embedded sensor in real cockpits. It provided a satisfying accuracy among 10 pilots out of 11 and more importantly, led to an accurate offline prediction of the flight phases using the group data of the first run, confirming some stability of this measure compared with other ones. In the prospect of real flight applications, future work will address this classification online using real-time eye event classification algorithms (Grindinger, 2006; Komogortsev & Karpov, 2013) and LDA analyses. This objective online information would help to understand the current state of the pilot and use it as a human-computer interface input (e.g., reducing the amount of information in overloaded situations or adapting it in underloaded ones) and/or at least, as a part of the flight recorder data for accident analyses (Peysakhovich, Lefrançois, Dehais, & Causse, 2018).

Limits and Future Work

Although the results presented here are encouraging for future real flight measurements, some limits have to be acknowledged. It is important to notice that the participants of the present study were relatively inexperienced. Hence, one could argue that the results obtained here may differ for more experienced pilots of general aviation (i.e., diminution of heart rate and visual pattern optimization). However, this limitation may be compensated by the fact that a plausible workload classifier could be trained on the data of a given pilot and used for this pilot whatever his or her experience, as shown in the work of Gateau and colleagues (2015) using functional NIRS. Another limit is the low number of flight phases considered that has arguably facilitated the classifier accuracy. Therefore, validation with several other phases has to be done before considering such an approach for real flights. Finally, the number of pilots involved in this study is small and has to be increased for stronger statistical validation. Thereby, future work involving several flight phases, a larger number of experienced pilots, and a real-time classification is planned and should provide the required evidence for real flight integration.

ACKNOWLEDGMENTS

The authors wish to thank all the pilots who participated in this study as well as the staff of Lasbordes aerodrome of Toulouse, France. This work was supported by the AXA research fund.

KEY POINTS

- To date, in-flight pilots' state is evaluated after the flight, mostly with debriefing subjective questionnaires.
- Most of the studies in aviation are made in flight simulators, limiting the interpretation to this environment. Those carried out in actual flights show some discrepancy about the usability of psychophysiological metrics to deduce pilots' state and have never considered the use of eye movements as a reliable pilots' state estimator.
- There is a need to develop online solutions that are straightforward to use and easy to implement in real cockpits to improve flight safety.
- We found that oculometric data from aircraft pilots could be used as a proxy in actual flight conditions. This approach allowed to discriminate pilots' states (as indexed by the flight phase) at the individual level. Because of drift issues, a cardiac-based classifier may need an additional online update to be more accurate.
- These results may pave the way to new potential applications for training, flight data recorder content, and human-machine interfaces.

ORCID ID

Sébastien Scannella https://orcid.org/0000-0001-9547-8303

Vsevolod Peysakhovich https://orcid.org/0000-0002-9791-4460

REFERENCES

Backs, R. W., & Walrath, L. C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, *23*(4), 243–254.

Blix, A., Stromme, S., & Ursin, H. (1974). Additional heart rate—An indicator of psychological activation. *Aerospace Medicine*, *45*(11), 1219–1222.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58–75.

Callan, D. E., Durantin, G., & Terzibas, C. (2015). Classification of single-trial auditory events using dry-wireless EEG during real and motion simulated flight. *Frontiers in Systems Neuroscience*, *9*, 11.

Casner, S. M., & Schooler, J. W. (2014). Thoughts in flight: Automation use and pilots' task-related and task-unrelated thought. *Human Factors*, *56*, 433–442.

Causse, M., & Matton, N. (2014). Using near infrared spectroscopy to detect mental overload in flight simulator. *Advances in Cognitive Engineering and Neuroergonomics*, *11*, 148.

Causse, M., Peysakhovich, V., & Fabre, E. F. (2016). High working memory load impairs language processing during a simulated piloting task: An erp and pupillometry study. *Frontiers in Human Neuroscience*, *10*, 240.

Christensen, J. C., Estepp, J. R., Wilson, G. F., & Russell, C. A. (2012). The effects of day-to-day variability of physiological data on operator functional state classification. *NeuroImage*, *59*(1), 57–63.

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, *250*, 126–136.

Dahlstrom, N., Nahlinder, S., Wilson, G. F., & Svensson, E. (2011). Recording of psychophysiological data during aerobatic training. *International Journal of Aviation Psychology*, *21*(2), 105–122.

Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., & Wickens, C. D. (2017). Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: A behavioral and eye tracking study. *International Journal of Aerospace Psychology*, *27*(1-2), 15–28.

Dehais, F., Causse, M., & Pastor, J. (2008). Embedded eye tracker in a real aircraft: New perspectives on pilot/aircraft interaction monitoring. In *Proceedings from the 3rd International Conference on Research in Air Transportation*. Fairfax, VA: Federal Aviation Administration.

Dehais, F., Causse, M., & Pastor, J. (2010). *Toward the definition of a pilot's physiological state vector through oculometry: A preliminary study in real flight conditions*. Retrieved from http://oatao.univ-toulouse.fr/11688/

Dehais, F., Causse, M., Vachon, F., Régis, N., Menant, E., & Tremblay, S. (2014). Failure to detect critical auditory alerts in the cockpit: Evidence for inattentional deafness. *Human Factors*, *56*, 631–644.

Dehais, F., Causse, M., Vachon, F., & Tremblay, S. (2012). Cognitive conflict in human-automation interactions: A psychophysiological study. *Applied Ergonomics*, *43*(3), 588–595.

Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., & Gateau, T. (2015). Automation surprise in aviation: Real-time solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2525–2534). New York, NY: ACM.

Dehais, F., Tessier, C., Christophe, L., & Reuzeau, F. (2010). *The perseveration syndrome in the pilot's activity: Guidelines and cognitive countermeasures*. New York, NY: Springer.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 271–285.

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., ... Pannasch, S. (2010). Saccadic peak velocity sensitivity to variations in mental workload. *Aviation, Space, and Environmental Medicine*, *81*(4), 413–417.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. New York, NY: Springer.

Durantin, G., Dehais, F., & Delorme, A. (2015). Characterization of mind wandering using fNIRS. *Frontiers in Systems Neuroscience*, *9*, 45.

Durantin, G., Gagnon, J.-F., Tremblay, S., & Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research*, *259*, 16–23.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PloS One*, *10*(3), e0121279.

Gouraud, J., Delorme, A., & Berberian, B. (2017). Autopilot, mind wandering, and the out of the loop performance problem. *Frontiers in Neuroscience*, *11*, 541.

Grindinger, T. (2006). *Eye movement analysis & prediction with the Kalman filter*. Unpublished thesis.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine*, *69*(4), 360–367.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Hughes, P. K., & Cole, B. L. (1998). The effect of attentional demand in eye movement behaviour when driving. In *Vision in Vehicles II* (pp. 221–230). London: Elseiver.

Jorna, P. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, *36*, 1043–1054.

Komogortsev, O. V., & Karpov, A. (2013). Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods*, *45*(1), 203–215.

Lee, Y.-H., & Liu, B.-S. (2003). Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation, Space, and Environmental Medicine*, *74*(10), 1078–1084.

Opmeer, C., & Krol, J. (1973). Towards an objective assessment of cockpit workload. i. Physiological variables during different flight phases. *Aerospace Medicine*, *44*(5), 527–532.

Peysakhovich, V., Lefrançois, O., Dehais, F., & Causse, M. (2018). The neuroergonomics of aircraft cockpits: The four stages of eye-tracking integration to enhance flight safety. *Safety*, *4*(1), 1–15.

Reynal, M., Rister, F., Scannella, S., Wickens, C., & Dehais, F. (2017). *Investigating pilot's decision making when facing an unstabilized approach: An eye-tracking study*. Retrieved from http://oatao.univ-toulouse.fr/18219/

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, *34*(2), 259–287.

Sauvet, F., Jouanin, J. C., Langrume, C., Van Beers, P., Papelier, Y., & Dussault, C. (2009). Heart rate variability in novice pilots during and after a multi-leg cross-country flight. *Aviation, Space, and Environmental Medicine*, *80*(10), 862–869.

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV-heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, *113*(1), 210–220.

Thomas, L. C., & Wickens, C. D. (2004). Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display. *Proceedings of the 48th Human Factors and Ergonomics Society Annual Meeting* (pp. 223–227). Santa Monica, CA: Human Factors and Ergonomics Society.

Togo, F., & Takahashi, M. (2009). Heart rate variability in occupational health: A systematic review. *Industrial Health*, *47*(6), 589–602.

Tokuda, S., Obinata, G., Palmer, E., & Chaparro, A. (2011). Estimation of mental workload using saccadic eye movements in a free-viewing task. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pp. 4523–4529). New York, NY: IEEE.

Veltman, J. (2002). A comparative study of psychophysiological reactions during simulator and real flight. *International Journal of Aviation Psychology*, *12*(1), 33–48.

Veltman, J., & Gaillard, A. (1993). Indices of mental workload in a complex task environment. *Neuropsychobiology*, *28*(1–2), 72–75.

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *International Journal of Aviation Psychology*, *19*(2), 182–199.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, *12*(1), 3–18.

Wilson, G., & Fisher, F. (1991). The use of cardiac and eye blink measures to determine flight segment in F4 crews. *Aviation, Space, and Environmental Medicine*, *62*(10), 959–962.

Sébastien Scannella is a research scientist at ISAE-SUPAERO, Toulouse, France, in the neuroergonomics and human factors team. He received his PhD in neuroscience from Université de Toulouse III–Paul Sabatier, France, in 2011.

Vsevolod Peysakhovich is a research scientist at ISAE-SUPAERO, Toulouse, France, in the neuroergonomics and human factors team. He received his PhD in computer science from Université de Toulouse III–Paul Sabatier, France, in 2016.

Florian Ehrig is a data analyst and digitalization engineer at Airbus, Toulouse, France. He received his MS in aerospace mechanics and avionics from ISAE-SUPAERO, Toulouse, France, in 2014.

Evelyne Lepron is CEO of EMOSCIENCES, SASU. She received her PhD in neuroscience from Université de Toulouse III–Paul Sabatier, France, in 2009.

Frédéric Dehais is a professor at ISAE-SUPAERO, Toulouse, France, in the neuroergonomics and human factors team. He defended his PhD in computer science in 2004 at ONERA (Office National d'Etude et de Recherche Aéronautique) on the topic of modeling cognitive conflict in pilots' activity.

Date received: October 9, 2017
Date accepted: June 4, 2018