







# Supporting Sensor-Based Usability Studies Using a Mobile App in Remotely Piloted Aircraft System

Antonio Esposito<sup>1</sup>(✉) , Giusy Danila Valenti<sup>1</sup> , Fabrizio Balducci<sup>2</sup> ,  
and Paolo Buono<sup>2</sup> 

<sup>1</sup> Università degli Studi di Enna Kore, Enna, Italy  
{antonio.esposito, giusy.valenti}@unikore.it

<sup>2</sup> Università degli Studi di Bari “Aldo Moro”, Bari, Italy  
{fabrizio.balducci, paolo.buono}@uniba.it

**Abstract.** Monitoring user workload during task performance is a relevant and widely investigated topic in the aviation field of study due to its associations with the level of safety and number of human errors. The current study aims at assessing the workload of pilots wearing sensors while performing typical fly operations. To this purpose, a mobile app able to record physiological measures while performing usability studies with multiple users, even remotely, is provided. Results coming from a preliminary test with three pilots reveal the usefulness of the app in the evaluation of the workload level for each participant and each task.

**Keywords:** Human factors · Physiological measures · Evaluation

## 1 Introduction and Related Work

Aviation safety is a complex domain that can be improved not only by taking into account the reliability of the aircraft and its systems, but also by considering and monitoring the crew performances. 75% of aircraft accidents are related to human errors that in most cases derive from the mental workload and fatigue of pilots and aircraft operators [18]. Human error can be a consequence of design flaws, inadequate training, incorrect procedures, old manuals and other technical, environmental and organizational factors [11].

The problem is increased in the context of *Remotely Piloted Aircraft Systems* (RPAS) that ranges from small devices, used for play purposes, to large aircraft, configured with sophisticated equipment and with extremely heterogeneous performances, dimensions and characteristics. A common element between these devices is represented by the absence of a pilot on board, since they are controlled by a ground operator from a remote station.

Aeronautics is one of the most interested field to issues related to Human Factors. Understanding the psycho-physical state of pilots and real-time monitoring of the mental workload during simulator training or in real flight can

improve the efficiency of a crew and the safety of air operations. Workload refers to the measurement of mental processing requests placed on a person during the execution of a task [15] and it can affect flight performance predisposing crews to mistakes [6, 20, 21]. The required amount of cognitive resources varies with the current activity. For example, operational phases like landing and take-off, which are the most critical [19] with the highest workload.

When humans have to perform physical and/or mental tasks, their cognitive state varies and many parameters such as fatigue, attention and memory must be monitored [16] measuring, for example, if person A is working harder than person B or whether task A requires more work than task B by defining practical and reasonable limits for the workload.

The contributions of this work are: i) a mobile app that interfaces with physiological sensors allowing to support user-based perform usability studies even remotely; ii) a pilot study that exploits the acquired data to evaluate the workload of drone pilots.

The paper is structured as follows: studies in the field of workload evaluation through behavioral, subjective and objective techniques are presented in Sect. 2; Sect. 3 introduces materials and methods to perform experimental evaluation through a mobile app and an electrocardiogram (ECG) sensor. In Sect. 4 there are results about the evaluation of the proposed approach; Sect. 5 provides conclusions and future works.

## 2 Evaluation Techniques

There are several ways to estimate mental workload: 1) Behavioral measurements: performance indices based on reaction times and the number of errors related to the task; 2) Subjective measurements: provided by the operator once the task is completed through specific questionnaires; 3) Physiological measurements: made with specific equipment that analyzes objective data like heart rate, respiratory rate, eye movements.

A study conducted by Alaimo et al. [2] analyzed the workload level during a flight mission performed using a simulator correlating Heart Rate Variability (HRV) biometric data with the subjective information collected through the NASA-TLX questionnaire. In particular, heart rhythm was used to determine the body's natural response to stressful situations, whereas subjective measures were used to analyze how the pilot perceives these workloads. From this perspective, more than one methodology should be considered for workload evaluation.

### 2.1 Behavioral Measures

Behavioral measures are directly related to the definition of performance provided by Paas et al. [22] that can be defined as the effectiveness in completing a particular task. Typical examples relating to the performance measurements can be considered: i) the detection of the reaction time or the time elapsed between

a stimulus and the execution of a response; ii) the pace to perform the activity; iii) the accuracy of the performance or the percentage of errors carried out during the tasks. In such a way, the workload can be assessed by analyzing the decrease in the subject performances. Nevertheless, according to Wilson et al. [27], the performance measures do not adequately reflect the level of workload experienced since the quality of behavioral measures does not always result in the operator's ability/inability to react adequately to the required task.

## 2.2 Subjective Measurements

Subjective measures are low cost and non-invasive assessment tools for evaluating workload [2, 9, 21].

*Inspection methods* involve expert evaluators and results depend on their own experience, sensitivity and personal preferences. For this reason, experts produce their evaluation individually and independently, then they meet and redact a final report that considers all of them. An example of an inspection technique is the 'Cognitive Walkthrough' which evaluates the task considering that users usually prefer to learn a system by using it rather than studying a manual [2, 9]. Answering questions for each action allows evaluators to analyze the problems that a user encounters in formulating objectives to perform a task, associating the correct action with the aim of carrying out actions.

*User-based* techniques allow to mitigate the subjectivity of the evaluations with experts. They consist of groups of users performing tasks in a controlled environment. A sample of users that is representative of the category to which the system is addressed is selected and everyone is asked to perform the same tasks separately. The expert observes and analyzes their behavior to understand if, where and why they have encountered difficulties. In the "Thinking aloud" technique users comment aloud thoughts while performing the tasks. Comments are related to the difficulties, and since the causes of such difficulties may not be evident, the observer must take notes, recording the situations in which the user is uncertain or commits some mistakes in order to review them later and identify issues and propose corrections.

Interviews and questionnaires are part of subjective techniques. They are usually employed in the aeronautical field to measure the user workload [17]. The *NASA-TLX (Task Load Index)*, [11, 24] the *Modified Cooper-Harper Workload Rating Scale* [26], and the *SWAT (Subjective Workload Assessment Technique)* are among the most common used questionnaires.

## 2.3 Physiological Measurements

Physiological measures, based on the detection of several physiological parameters, can be considered as a very powerful technique for assessing workload, and they are widely applied in numerous studies [3, 7, 25]. Using physiological measures in the aviation field of study is advantageous for several reasons. First, these kinds of measures, which are also tested and used in the medical sector,

have a high level of validity and reliability. In addition, they provide an objective assessment since their evaluation does not depend on subjective perceptions. Moreover, the detection of physiological parameters can allow a real-time workload assessment, leading to continuous monitoring of variations in the amount of physical and mental effort experienced while performing a task [27]. The electroencephalogram (EEG), electrocardiogram (ECG), pupil size, and the eye movement are very useful methods for measuring workload, and they seem to overcome some of the shortcomings of both subjective and behavioral measurements [8] and have been successfully employed to evaluate human aspects like emotions [4].

From this perspective, the shift from low to high workload can be evaluated in terms of changes of the activity of the Autonomous Nervous System (ANS), which can be linked to specific physiological responses. Among the different physiological measures used in this specific sector, indexes based on Heart Rate Variability (HRV) measurement are very popular thanks to their high level of sensitivity and efficiency in assessing changes in mental workload [2, 13, 21]. [10]. HRV is inversely associated with workload: individuals experiencing high levels of mental effort tend to show a decrease of HRV due to the activation of Sympathetic Nervous System (SNS) and/or to the Parasympathetic (PNS) withdrawal [23].

The indexes based on HRV measurements can be distinguished into different categories, and the most common are based on: i) the time domain analysis, ii) the frequency domain (or power spectral density) analysis, and iii) the non-linear indexes analysis. The time-domain parameters quantify the amount of variability in measurements of the Interbeat Interval (IBI), meant as the time period between two consecutive heartbeats. Some of the most common time-domain metrics are the Standard Deviation of NN intervals (SDNN) and the Root mean square of successive RR interval differences (RMSSD). The frequency domain indexes involve defining ECG waves as different spectral components through power spectral density and they can be divided into four frequency bands: HF (High Frequency) frequencies between 0.15 and 0.4 Hz; LF (Low Frequency) frequencies between 0.04 and 0.15 Hz; VLF (Very Low Frequency) frequencies between 0.003 and 0.04 Hz; ULF (Ultra Low Frequency) the band falls below 0.003 Hz. Given the very low frequency of the oscillations, the ULF band's contribution can only be appreciated in 24-hour acquisitions. Nonlinear indexes used in HRV analysis include Poincaré plot, detrended fluctuation analysis, approximate entropy, and sample entropy calculation [12].

### 3 Materials and Methods

We developed a mobile app to support usability experts to carry out evaluation studies. The app allows to evaluate the subject's mental load during the performance of real tasks by measuring physiological data produced by specific devices. In this work ECG data detected by a Polar H10 sensor have been used.

### 3.1 The ECG Sensor

The Polar H10 is equipped with an elastic strap with integrated electrodes. It records data relating to inter-beat times and Pulse-to-Pulse Interval (PPI) like ECG, with a sample rate 130 Hz and an impedance of  $2M\Omega$  in a range of  $\pm 20,000 \mu V$ . The signal quality of RR intervals (distance between two R peaks of ECG wave) of a Holter device and the Polar H10 chest strap at rest and during exercise both measurement systems are valid for RR intervals detection [14]. The sensor uses Bluetooth (BLE) or Ant + technologies to communicate and allows the connection to multiple devices simultaneously. Polar software libraries allows developers to extract live data directly from the sensors, in order to acquire in real-time *HR data* (bpm), *RR interval* data (1/1024 format), *electrocardiogram data* ( $\mu V$ ), with timestamp, and *record HR and RR data* in the device internal memory using a sample time of 1 s.

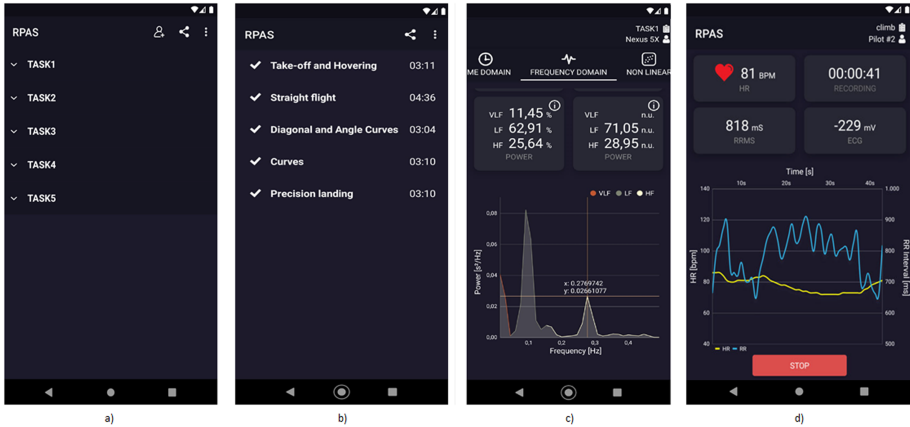
### 3.2 The Mobile App

The app connects to the Polar H10 sensor through the Bluetooth BLE connection supporting test sessions with two non-exclusive configurations: *On-site*: users take the experimental device one at a time wearing the sensor while carrying out their tasks. The evaluator manages different participant sessions through the app; *Remote*: the information about the tasks to perform are sent to the participant's device who wears the sensor, so that it can record ECG data. Data are sent to the evaluator upon completion of all the tasks in a session.

Data coming from local users are combined with the remote ones, flowing neatly into the experimenter's device and thus building the global dataset. The final results is the performing of a hybrid user-test session. When a configuration file is opened on a device, the tasks to be performed are loaded into the app and the device is ready to start the test. Once all the tasks have been carried out, the file containing the measurements is saved and, in case of a remote user, is sent to the evaluator. The data of the session and tasks are stored in an SQLite DBMS and the app exports data in raw format as .csv file and as human-readable tables. Everything is compressed in a .zip file that the evaluator can easily manage.

The initial screen of the app shows the list of the sessions. In sessions created in the device, the evaluator can add participants. The app displays the list of subjects (local and remote) who participated in a specific session and the time to complete each task. Each participant sees only personal tasks and those already completed (Fig. 1a), displayed with a check on the left of the task name and the duration time taken to complete it on the right (Fig. 1b). Figure 1c shows an example of results in the frequency domain, in addition to the values (top), the Power Spectral Density of the signal (bottom) is given.

To create a new experimental session the evaluator specifies the name of the test and the number of tasks to perform, then it is possible to rename them or keep the automatic names as  $TASK_i$  ( $i = 1, \dots, n$ ). In order to carry out a task where data from the sensor are recorded, a message shows the status of the connection and, once established, values are shown in real-time with a graph



**Fig. 1.** Four screenshots of the app showing, respectively a) the task list, b) the same list after the task have been renamed by the evaluator, c) the sensor details in the frequency domain and d) the real-time HR and RR data produced by the sensor.

that depicts the HR values in bpm on the left Y axis while on the right Y axis are the RR values in ms (Fig. 1d). At the end of a task, by pressing the “STOP” button on the panel the file is saved.

## 4 Experiments and Results

A user study to examine the app effectiveness on the HRV acquired measurements was carried out. Three drone pilots, all having an official drone pilot’s license, were engaged for a flight mission based on five different tasks. In the following section, the drone missions success rates are presented, and successively, an example of the physiological measurements for the evaluation of workload during two different flight phases is described.

### 4.1 Drones Flight Experiment

The evaluation of the usability of piloting systems has been performed exploiting two types of drones: *DJI Fly* app for the *DJI Mavic Mini* drone and *FIMI Navi* app for the *Xiaomi Fimi X8 SE* drone. After having interviews with the subjects regarding their previous experience with drones, five tasks with different complexity were designed to stress specific functions. The tests were conducted individually with three participants, one at a time, and were carried out in the open field. Each pilot used a drone with active stabilizer and Global Positioning System (GPS). The task to execute were:

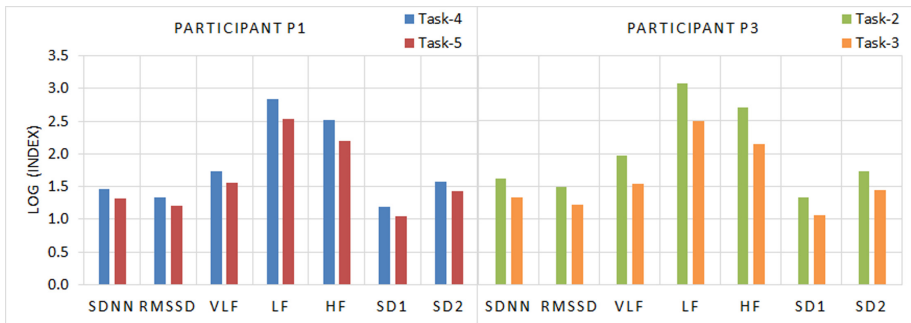
1. *Take-off and Hovering*: take altitude up to about 10 m for a few seconds and then descend

2. *Straight flight*: proceed straight at constant altitude for 30 m and go back to the starting position by turning (use of the inverted controls)
3. *Diagonal and Angle curves*: from a low altitude made a diagonal climb of  $45^\circ$ , draw a rectangle in flight and then make a diagonal descent of  $45^\circ$
4. *Curves*: fly to the agreed point, circle around it and return to the starting point
5. *Precision Landing*: land the drone at a designated point

**Table 1.** Results of the evaluation. For each on the five tasks, the colored cells report the performance results for a participant  $P_i$  where S indicates *Success*, F indicates *Failure* and P indicates *Partial* success in the task correct execution.

	Task-1	Task-2	Task-3	Task-4	Task-5
P1	S	S	S	S	F
P2	S	S	S	S	P
P3	S	S	F	S	S

As visible in Table 1, while most of the tasks were successfully completed, two failures and one partial success were recorded. Partial success was attributed in the event that unforeseen external factors influenced the difficulty.



**Fig. 2.** HRV indexes between task 4 and task 5.

## 4.2 Pilot Workload Evaluation

During the drone flight phases, the heart rate variability of the pilots was recorded and monitored by the app. The measurements of HRV indexes can usefully estimate the workload levels associated with the task's difficulty. Indeed, the pilot's emotional state influences the HRV when the participant is not able

to complete the task successfully. To analyze this effect and compare a successful task with a failure task, participants P1 and P3 are considered. The results in Fig. 2 for Task-4 and Task-5, and Task-2 and Task-3 are given respectively for P1 and P3. Results are reported as logarithmic values in the y-scale to compare the HRV indexes on the same amplitude scale. The HRV indexes taken into account are inversely associated with the levels of workload (for more details the reader could refer to [1]). Concerning the results, it is worth noting that for all the HRV indexes, Task 5-is associated with lower values than Task-4; and the same effect is found for Task-3 and Task-2. These results can be read as an increase in the workload level while accomplishing the task or increasing mental workload levels due to the failure. Moreover, in particular, the workload associated with the precision landing flight phase is higher than other maneuvers as reported in [2,21]; additionally, the level flight can be considered as a lower demand maneuver and therefore associated with higher HRV values as shows in Fig. 2 for the Task-2.

## 5 Conclusion and Future Work

This paper is focused on the support to evaluators in conducting usability study involving participants that wear physiological sensors. An app that connects to the sensors, collects the physiological data, makes the computation, shares the data to the evaluator and creates a basic report that summarizes the main collected data has been developed and briefly presented. We conducted a test with three pilots that wore a Polar H10 sensor and performed five typical tasks having different difficulty.

The use of an app to monitor physiological data has several advantages: the evaluator is leveraged in taking the task execution times; the data collected from the sensors are computed and stored directly by the app; the data of the test are easily shared between the user and the evaluator, also in the case that they are not co-located; the setting works well also in real scenarios and does not require a lab. Indeed, the tests with pilots presented in this paper were performed in an open field. Several apps capable of manipulating the sensor data are available but, to the best of our knowledge, none also directly connect to the sensor, support usability studies, store the data and show both real-time information and the final report of the session.

We plan to integrate new sensors to increase the subjective data, such as ECG and Galvanic Skin Response. We also plan to add subjective evaluations in the app (e.g. questionnaires). The integration the two types of workload assessment will help better estimating the human factors and associated practices for the workload in aircraft crews. The results presented in this work may also be useful to the evaluation and improvement of Smart Interactive Experiences [5] (in the Cultural Heritage domain), and in general when the subject is at the center of experience and IoT sensors help collecting data.

**Acknowledgments.** This work is partially supported by the projects “Integrazione dei Sistemi Aeromobili a Pilotaggio Remoto nello spazio aereo non segregato per



servizi” (RPASinAir ARS01.00820 CUPJ66C18000460005), funded by Italian Ministry of Education, Universities and Research (MIUR) and “Gestione di oggetti intelligenti per migliorare le esperienze di visita di siti di interesse culturale” funded by the Apulia Region under the program Research for Innovation (REFIN) POR Puglia FESR FSE 2014–2020. The authors thank Edilio Formica for his help in the app implementation.

## References

1. Alaimo, A., Esposito, A., Orlando, C.: Cockpit pilot warning system: a preliminary study. In: 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI), pp. 1–4 (2018). <https://doi.org/10.1109/RTSI.2018.8548518>
2. Alaimo, A., Esposito, A., Orlando, C., Simoncini, A.: Aircraft pilots workload analysis: heart rate variability objective measures and Nasa-task load index subjective evaluation. *Aerospace* **7**(9), 137 (2020). <https://doi.org/10.3390/aerospace7090137>
3. Baevsky, R.M., Chernikova, A.G.: Heart rate variability analysis: physiological foundations and main methods. *Cardiometry* (10) (2017)
4. Balducci, F., Grana, C., Cucchiara, R.: Classification of affective data to evaluate the level design in a role-playing videogame. In: 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), pp. 1–8 (2015). <https://doi.org/10.1109/VIS-GAMES.2015.7295766>
5. Balducci, F., Buono, P., Desolda, G., Impedovo, D., Piccinno, A.: Improving smart interactive experiences in cultural heritage through pattern recognition techniques. *Pattern Recognit. Lett.* **131**, 142–149 (2020). <https://doi.org/10.1016/j.patrec.2019.12.011>. <http://www.sciencedirect.com/science/article/pii/S0167865519303745>
6. Boff, K.R., Kaufman, L., Thomas, J.P.: *Handbook of perception and human performance* (1986)
7. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **44**, 58–75 (2014)
8. Brookhuis, K.A., De Waard, D.: Monitoring drivers’ mental workload in driving simulators using physiological measures. *Accid. Anal. Prev.* **42**(3), 898–903 (2010)
9. Cao, X., et al.: Heart rate variability and performance of commercial airline pilots during flight simulations. *Int. J. Environ. Res. Public Health* **16**(2), 237 (2019)
10. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental workload alters heart rate variability, lowering non-linear dynamics. *Front. Physiol.* **10**, 565 (2019)
11. Dumitru, I.M., Boşcoianu, M.: Human factors contribution to aviation safety. *Sci. Res. Educ. Air Force-AFASES* **2015**(1), 49–53 (2015)
12. Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing: Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* **93**(5), 1043–1065 (1996)
13. Fuentes-García, J.P., Clemente-Suárez, V.J., Marazuela-Martínez, M.Á., Tornero-Aguilera, J.F., Villafaina, S.: Impact of real and simulated flights on psychophysiological response of military pilots. *Int. J. Environ. Res. Public Health* **18**(2), 787 (2021)
14. Gilgen-Ammann, R., Schweizer, T., Wyss, T.: RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *Eur. J. Appl. Physiol.* **119**(7), 1525–1532 (2019). <https://doi.org/10.1007/s00421-019-04142-5>

15. Gopher, D., Donchin, E.: Workload: an examination of the concept (1986)
16. Hancock, P.A., Matthews, G.: Workload and performance: associations, insensitivities, and dissociations. *Hum. Factors* **61**(3), 374–392 (2019). <https://doi.org/10.1177/0018720818809590>
17. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Human Mental Workload*, *Advances in Psychology*, North-Holland, vol. 52, pp. 139–183 (1988). [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
18. Kharoufah, H., Murray, J., Baxter, G., Wild, G.: A review of human factors causations in commercial air transport accidents and incidents: from 2000–2016. *Prog. Aerosp. Sci.* **99**, 1–13 (2018). <https://doi.org/10.1016/j.paerosci.2018.03.002>
19. Lee, Y.H., Liu, B.S.: Inflight workload assessment: comparison of subjective and physiological measurements. *Aviat. Space Environ. Med.* **74**, 1078–84 (2003)
20. Liu, J., Gardi, A., Ramasamy, S., Lim, Y., Sabatini, R.: Cognitive pilot-aircraft interface for single-pilot operations. *Knowl.-Based Syst.* **112**, 37–53 (2016). <https://doi.org/10.1016/j.knosys.2016.08.031>
21. Mansikka, H., Virtanen, K., Harris, D.: Comparison of NASA-TLX scale, modified cooper-harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics* **62**, 1–22 (2018). <https://doi.org/10.1080/00140139.2018.1471159>
22. Paas, F.G.W.C., Merriënboer, J.J.G.V.: The efficiency of instructional conditions: an approach to combine mental effort and performance measures. *Hum. Factors* **35**(4), 737–743 (1993). <https://doi.org/10.1177/001872089303500412>
23. Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A., Van Huffel, S.: Instantaneous changes in heart rate regulation due to mental load in simulated office work. *Eur. J. Appl. Physiol.* **111**(7), 1497–1505 (2011). <https://doi.org/10.1007/s00421-010-1776-0>
24. Valdehita, S., Ramiro, E., García, J., Puente, J.: Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* **53**, 61–86 (2004). <https://doi.org/10.1111/j.1464-0597.2004.00161.x>
25. Wanyan, X., Zhuang, D., Zhang, H.: Improving pilot mental workload evaluation with combined measures. *Bio-Med. Mater. Eng.* **24**(6), 2283–2290 (2014)
26. Wierwille, W.W., Casali, J.G.: A validated rating scale for global mental workload measurement applications. In: *Proceedings of the Human Factors Society Annual Meeting*, vol. 27, no. 2, pp. 129–133 (1983). <https://doi.org/10.1177/154193128302700203>
27. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Hum. Factors* **45**(4), 635–644 (2003)