Toward the Use of Pupillary Responses for Pilot Selection

Nadine Matton, Pierre-Vincent Paubel, Sebastien Puma

Title : Towards the use of pupillary responses for pilot selection

Nadine Matton

ENAC, University of Toulouse, France.

CLLE, University of Toulouse, CNRS, France

Pierre-Vincent Paubel

CLLE, University of Toulouse, CNRS, France

Sébastien Puma

Paragraphe laboratory, University of Cergy Pontoise, EA 349, France

CLLE, University of Toulouse, CNRS, France

Précis/Short abstract: Task-induced mental workload of pilot students was assessed at the beginning of their training on a laboratory task. Their two year flight training outcome was associated with characteristic pupil size variations, discriminating the most and less proficient students. Results could be interpreted in line with the neural efficiency hypothesis.

Acknowledgments**:** The authors would like to acknowledge Jean-Baptiste Gervais for his work placement report that helped us write the introduction of the paper and Camille Jeunet for her helpful comments on the manuscript.

Correspondence concerning this article should addressed to Nadine Matton, ENAC Research Lab, 7 avenue Edouard Belin 31055 Toulouse, France. E-mail: nadine.matton@enac.fr

Abstract

Objective: For selection purposes, it seems important to assess the level of the mental resources invested in order to perform a demanding task. In this study, we investigated the potential of pupil size measurement to discriminate the most proficient pilot students from the less proficient.

Background: Cognitive workload is known to influence learning outcome. More specifically, cognitive difficulties observed during pilot training are often related to a lack of efficient mental workload management.

Method: Twenty pilot students performed a laboratory multitasking scenario, composed of several stages with increasing workload, while their pupil size was recorded. Two groups of pilot students were contrasted according to the outcome after two years of training, namely High and Medium success.

Results: Our findings suggested that task-evoked pupil size measurements could be a promising predictor of flight training difficulties during the two year training. Indeed, High success pilot students showed greater pupil size changes from low-load to high-load stages of the multitasking scenario than Medium success pilot students. Moreover, average pupil diameters at the low-load stage were smallest for the High success pilot students.

Conclusion: These results were interpreted within the neural efficiency hypothesis framework, the most proficient pilot students using supposedly their mental resources more efficiently.

Application: These findings might open a new way of managing selection processes complemented with ocular measurements. More specifically, pupil size measurement could enable to identify applicants with greater chances of success during pilot training.

Keywords: Mental workload; eye behaviour; individual differences; pilot training; neural efficiency

Introduction

Pilot selection is carried out by most of civil or military pilot training schools or by airlines (e.g., Carretta & Ree, 1994; Huelmann & Oubaid, 2004; Martinussen & Torjussen, 1998). Traditionally, pilot selection processes use predictors derived from performances in cognitive ability tests, although these measures proved to be only moderately efficient (Damos, 1993). To the best of our knowledge, no current selection system uses physiological data collected while applicants perform the cognitive ability tests.

Pupil dilation is a non-invasive and objective measurement related to the mental workload induced by performing a task (e.g., Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Kang, Hueffer & Wheatley, 2014). Although there is still a debate whether the pupillary response is related to emotional arousal or to cognitive effort (e.g., Bradley, Miccoli, Escrig & Lang, 2008; Chen & Epps, 2013; Partala & Suraka, 2003), for the sake of our study, we retain that the pupillary response may be related to mental workload, i.e. emotional and/or cognitive load. Indeed, pilot training encompasses both cognitive and emotional factors. The rationale of this paper is that pupil dilation could complement behavioural measurements during pilot selection, as it enables the assessment of the amount of mental effort involved while performing a complex task. With equivalent performances of applicants, the objective would be to discriminate between those who devote only little mental effort from those who devote a large amount of effort. Indeed, given equal performance between applicants, selection of applicants who exert less cognitive effort would be preferable, because they are supposed to use their mental resources more efficiently. In other words, this paper aimed at exploring the predictive validity of pupillary responses during a selection process for pilot training.

Individual differences in pupillary responses

Early empirical research about task-induced variations in pupil diameter found that pupil diameter increased with the difficulty of a memory load task (Hess & Polt, 1964; Kahneman & Beatty, 1966). Since these first research works, task-induced variations in pupil diameter have been widely studied with some research focusing on individual differences. Findings from studies about individual differences in task-evoked pupillary responses are quite inconsistent.

A first group of studies has found that low performers showed greater pupil dilations (e.g., Ahern & Beatty, 1979; Heitz, Schrock, Payne and Engle, 2008 in the condition without incentives; Jainta & Baccino, 2010), suggesting that low performers needed to recruit more resources to complete the tasks, particularly when difficulty increased. In the present paper, the term resources is referring both to working memory and to controlled attention. Ahern and Beatty (1979) have shown that low performers at the scholastic aptitude test (used for college admissions in the United States) had larger task-induced pupillary responses than high performers during simple tasks (multiplications, digit span, sentence comprehension). Heitz et al. (2008) have measured pupillary responses during a reading span task and compared high *versus* low performers during an operation span task. The task-evoked pupillary responses during the recall phase of the reading span task revealed that low-span individuals had greater pupil diameter changes from their baseline than high-span individuals, specifically when no incentive was provided. Similarly, Jainta and Baccino (2010) have observed larger pupil increases for individuals who had made the most errors on a multiplication task compared to those who had made fewer errors.

On the contrary, a second group of studies has highlighted greater pupil dilations for high performers (e.g., Tsukhara, Harrison and Engle, 2016 for the high load trials; van der Meer et al., 2010 with the geometry analogy task) suggesting that high performers have more spare resources, specifically when the tasks are difficult. Van der Meer et al. (2010) found that on a

difficult task (geometric analogies), individuals with high fluid intelligence scores (assessed through the Advanced Progressive Matrices test) had greater pupil dilations than individuals with average APM scores. Recently, Tsukhara et al. (2016) found that high performers on working memory tasks (and also with high performers on fluid intelligence tasks) had higher pupil dilations on a simple memory task than low performers, specifically when the task became very demanding (from 9 to 12 letters to be memorized).

Finally, a third group of experiments found no differences in pupil responses for high and low performers (Heitz et al., 2008 in the condition with incentives; van der Meer et al., 2010 with the choice reaction time task). Importantly, all these studies varied in the way they identified individual differences (e.g., with a pre-experimental assessment or based on task performances) and in the nature of the task performed during pupil measurement (e.g., simple multiplications or difficult analogical reasoning).

Depending on the specificity of the task and of the context, high performers may be characterized by a greater efficiency in the way they use their resources or by a greater level of spare resources to cope with a demanding situation. High performers may both have more resources available, specifically with unknown tasks or at the beginning of the training and may be more efficient in the way they use them after practicing these tasks. As a consequence, for difficult tasks or conditions, the pupil size of high performers should be higher than for low performers and for easy tasks or conditions, the pupil size of high performers should be smaller than for low performers. This hypothesis is in line with the *neural efficiency hypothesis* of intelligence (Neubauer & Fink, 2009) and specifically with its interaction with task difficulty. Indeed, many studies based on brain measurements have found that more proficient individuals displayed lower brain activations when performing low to moderate difficulty tasks or after sufficient practice of complex tasks. On the other hand, they invested more brain resources when the task was very difficult, compared to less proficient individuals (e.g., Di Domenico,

Rodrigo, Ayaz, Fournier & Ruocco, 2015; Doppelmayr et al., 2015; Dunst et al., 2014; Lipp et al., 2012; Puma, Matton, Paubel, Raufaste & El-Yagoubi, 2018). Therefore, in a context of selection, it would be helpful to identify those individuals who are likely to invest spare resources when the task becomes more demanding. The objective of the present study is to assess the predictive validity of task-evoked pupillary responses for a highly demanding training, namely airline pilot training.

Pilot training

Pilot training is a highly complex learning process that requires trainees to acquire a large amount of knowledge (procedures, rules, aircraft laws, meteorology, etc.) and to develop many skills. For a student with no flying experience, airline pilot training lasts 2.5 to 3 years and is composed of theoretical and practical training. The practical training is composed of flying hours with a flight instructor and simulator flights. In case of learning difficulties, the attribution of additional flying hours or the exclusion from the training is decided by the training organization. A frequently reported difficulty of pilot trainees during practical training is the lack of spare mental resources in order to manage the flight. In a similar context, military flight instructors reported that the most frequent difficulties of pilot trainees were related to "attention control" and inability to deal with the "load of flight" (Gopher, Weil & Bareket, 1994, p.388). Indeed, managing a flight is a challenge for pilot trainees, as they have to manage the attitude of the aircraft (pitch, roll and heading), anticipate their trajectory, communicate with the other actors (like air traffic controllers or their flight instructor) and manage the systems in the cockpit, all this under time pressure. Consequently, the pilot trainee might feel overloaded. The question that arises is whether this lack of spare mental resources during flight training could be associated with a lack of mental resources while performing a demanding laboratory task. Many pilot selection processes include a multitasking test in order to assess the ability of the applicant to perform several tasks concurrently. A meta-analysis of the predictive validity of

performance scores derived from such tests has revealed a significant correlation with training outcome. Still, the global effect remains small ($r = .23$, Damos, 1993). This means that obtaining a high performance on such a test is not necessarily linked with the pilot's training success. Indeed, a high performance may be obtained with a more or less high level of effort (Wickens, 2002).

Overview of the study

The present paper focused on the capacity of individual differences in task-evoked pupillary responses to predict learning individual differences in an ecologically valid environment. More precisely, we studied the relationship between pupillary responses on a demanding laboratory task and the outcome of airline pilot training. The rationale of the study is that if the differences in pupillary responses in a laboratory task are meaningful and ecologically valid, then they should be associated with differences in training outcome.

In the present study, pupillary responses of pilot trainees were recorded while they were performing a psychomotor multitasking scenario before the beginning of their practical training. This task consisted in six successive stages with one to four subtasks to be performed concurrently. Therefore, the mental workload induced by this task was supposed to increase from low-load to high-load. Then, two groups of pilot trainees were identified after practical flight training completion (1.5 to 2 years after the recording of the pupillary responses), (i) trainees who successfully completed the training without notable difficulties and (ii) trainees who completed the training with more notable difficulties. The aim of the study was to identify whether these two groups of trainees would have different patterns of pupillary responses measured during the multitasking laboratory task.

Hypotheses

Following the neural efficiency hypothesis and, more specifically, its interaction with task difficulty, the following hypotheses could be derived:

H1: Given their higher level of spare mental resources, the most proficient pilot students should show greater pupil size changes from low-load to high-load stages when performing the complex psychomotor multitasking scenario, in comparison to less proficient pilot students.

H2: Given their higher mental efficiency, the most proficient pilot students should show lower pupil sizes during the lowest load stage of the psychomotor task, in comparison to less proficient pilot students.

Method

Participants

Twenty pilot students (18 men, 2 women) from the French civil aviation university volunteered and received a movie ticket after the experiment. They were all aged between 18 and 23 years ($M = 20.9$, S.D. = 1.3). They had been selected after a multiple stage selection process comprising scientific knowledge tests, English tests, cognitive ability tests, group exercises and individual interviews. Moreover, all of them came from scientific preparatory classes for competitive admission to elite universities. The Priority Management Task used in the present paper was the same as the one used during the cognitive ability tests stage of the selection. Thus, all the participants performed sufficiently well at this task to be selected. At the time of the experiment, all the participants were following their pilot theoretical training and had not started the practical phase of the training yet. This study was conducted in accordance with the principles expressed in the Declaration of Helsinki. Informed consent was obtained from each participant.

Procedure

Firstly, the pilot students performed a psychomotor multitasking scenario individually. Participants sat in front of a 19-inch, 1024x768 resolution computer screen and interacted through two joysticks and a keyboard. Pupil size was measured using an EyeLink 1000 desktop eye tracker (SR Research Ltd., Mississauga, Ontario, Canada). This eye-tracker possesses a spatial accuracy greater than 0.5°, and a 0.01° spatial resolution. The sampling rate was set to 1000Hz. A chin and forehead rest was used to maintain these distances and to avoid head movements. All eye-tracking data was extracted using the SR Research default algorithm. The experiment took place in a quiet simulation room. The volunteers were seated on a comfortable chair. The eye tracker camera was placed at a distance of 20 cm from the screen and the eye camera was at a distance of 60 cm from the screen. The lighting of the room was maintained constant and the luminance was controlled and remained constant throughout the experiment (212 lx as measured *a posteriori* with an analogical luxmeter Extech 401 025). Before starting the experimental phase, participants performed a short calibration phase in order to adjust the eye tracker. Participants were then told to perform the task in strict accordance with the instructions provided on the screen.

The multitasking scenario, labelled the Priority Management Task (see Figure 1), was organized in six successive four-minute stages except the first stage that lasted three minutes (see Matton, Paubel, Cegarra & Raufaste, 2016, for a detailed description of this task and the computation of the performance). To summarize, participants had to complete four concurrent subtasks that were successively added (from only one subtask at the first stage S1 to four subtasks at the fourth stage S4): Monitoring, Tracking, Detecting and Calculating. The 'Monitoring task' consisted in maintaining the levels of four gauges within an interval by using a first joystick. Every 15s one of the gauges deviated from its position at a speed of 10 to 70 pixels per second. To maintain the level of the gauge in the target interval, the participant had

to press on several buttons in order to select the right gauge and then to adjust the gauge to the desired value with the joystick. The 'Tracking task' consisted in keeping a cross positioned in a moving circle of 50 pixels diameter, in an area bordered by a large circle of 300 pixels diameter, through the second joystick. The circle moved every 15s at a speed of 3 to 12 pixels per second. The 'Detecting task' consisted in detecting the presence of three target letters (which varied from stage to stage) in a block of nine letters. Participants had to press one of nine keyboard keys as quickly as possible when a target letter appeared in the corresponding zone. A new block of letters was presented every 15s. The 'Calculating task' consisted of simple arithmetic problems (e.g., deducing a distance from given speed and time). The participants had to type the numeric answer as quickly as possible. Whether an answer had been given or not, a new problem was presented every 15s. At each stage, instructions indicated that each subtask was equally important and the number of concurrent subtasks increased from stage 1 (Monitoring subtask only) to stage 4 (all four subtasks concurrently). Two final stages, S5 and S6, varied the assigned priorities of the four subtasks. Subtasks were appended in the same order for all participants. All the characteristics of the subtasks were exactly the same for each participant. Each subtask required an action from the participant every 15s at the same time. Thus, mental workload was supposed to increase from stages S1 to S4 with the addition of a new subtask at each stage. Performance indices were computed in 100ms steps. For each of the subtasks, this ranged from 0 (worst) to 100 (best). Performance on Tracking was proportional to the distance between the cross and the moving circle. It was given a score of 100 whenever the cross remained within the moving circle. Depending on the speed of the moving circle, performance could decrease at a rate of 0 to 20% per second. Performance on Monitoring was proportional to the maximum distance between the four gauge levels and their corresponding target intervals. Moreover, the performance was given a score of zero whenever one gauge level went beyond a 60% tolerance interval. Depending on the speed of the gauge, performance could

decrease at a rate of 0% to 33% per second. Performance on Detecting and Calculating followed the same principle: Performance was given a score of 100 when the block of letters or the arithmetic problem was presented. Then the performance level gradually fell at a rate of 6.67 per second until the correct answer was keyed in. If a wrong answer was supplied, performance was more substantially decreased. At each stage, a global performance was computed by averaging all subtask performances. The instantaneous performance level of each subtask was displayed through a corresponding gauge in the top center of the screen. When performance on one of the subtasks dropped below a 10% performance threshold, the global performance was set to 0, in order to avoid participants neglecting one or several subtasks.

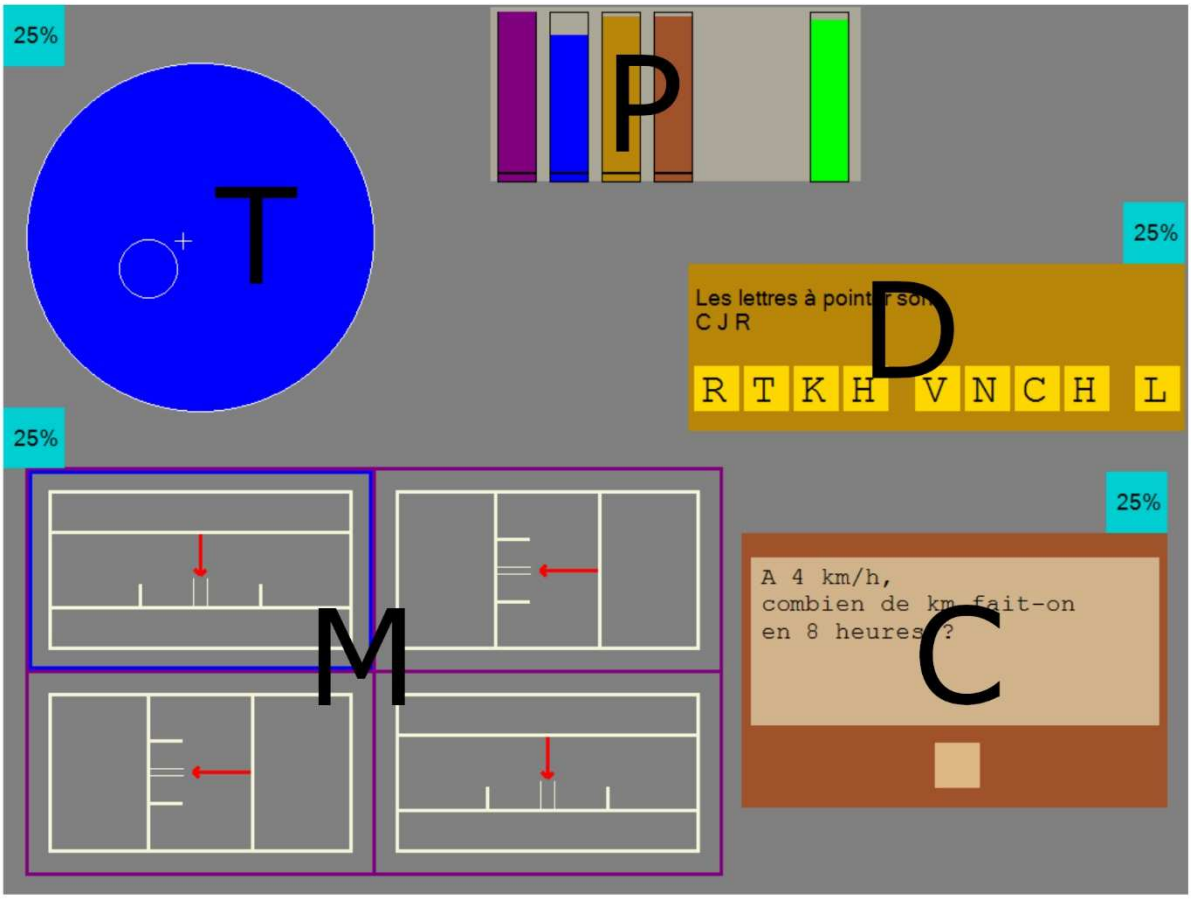

Figure 1. Screenshot of stage S4 of the Priority Management Task. The letters (*T* stands for tracking, *M* for monitoring, *D* for detecting and *C* for calculating subtask) have been

superimposed here for the reader but were not displayed during the task. *P* stands for performance gauges.

Secondly, we collected data of the training outcome of these pilot students after completion of practical training, i.e. two years later. We focused on the number of additional practical training hours they had received and we also collected some qualitative information regarding the training difficulties encountered. Additional practical training hours are requested by flight instructors and decided by the training organization. Given the cost of flight hours, a high number of additional flight hours is only given in case of substantive learning difficulties. Pilot students were divided into two groups, Medium ($n = 6$) and High ($n = 14$) levels, according to whether they had experienced such learning difficulties or not (the threshold was set at three hours of remedial training, see results for more details). Among the 20 pilot students that took part to this study, none of them was excluded from the training, thus no Low level group was identified. The "Low" term was kept unused to account for potential training failure during potential replications. Pilot students took the Priority Management Task with eye tracking recording during theoretical training, that means before flight training and none of them had yet received any remedial training. Therefore, none of the events that led to pilot categorization could have any influence on pupillometry or performance on the psychomotor task. Moreover, during the training completion, pilot instructors were unaware of the experiment being run and prescribed complementary flight hours as they usually did to students unable to achieve the mandatory level for final certification of air pilot training. The flight instructors discussed the threshold to consider medium or high performer students without consideration of the experiment.

Analyses

Fixations were extracted by the Eyelink standard algorithm and pupil sizes were only analysed for fixations. Thus, no filtering for blinks nor saccades was needed. Pupil sizes were extracted with the Eyelink arbitrary unit and then converted into millimeters according to the formula given in the Eyelink manual (i.e., pupil size (mm) = square root(pupil size(arbitrary unit))/10). Firstly, we computed average pupil size per one minute periods for each stage (three blocks for the three minute stage S1 and four blocks for all other four minute stages) that were used as repeated measurements within a stage. As the first stage (S1) consisted in performing only one simple monitoring task, this stage was identified as the lowest load stage. As we were interested in assessing the increase in use of mental resources when the workload rose, we decided to compute differences between each block average pupil size of the subsequent stages (S2 to S6) and the mean of the pupil size of the lowest load stage (S1) and called it *pupil size variation from S1*. In other words, we further analysed the following mean differences: (S2-S1), (S3-S1), (S4-S1), (S5-S1) and (S6-S1). Thus, we did not subtract any resting baseline value but the mean pupil size at the lowest-load stage. Beatty and Lucero-Wagoner (2000, p.148) recommended the baseline value to be subtracted from peak dilation or average pupil diameter. We preferred average pupil diameter because of its higher robustness. A 2 x 6 mixed analysis of variance was then conducted on the pupil size changes from stage S1 with a between-subject factor (Training outcome, High or Medium) and a within-subject factor (Stage, from S1 to S6). Secondly, raw pupil sizes were analysed as they provided further explanation on the dynamics observed when considering pupil size variations.

Statistical analyses. All analyses were computed using *R* (R Core Team, 2018) and all ANOVAs were conducted with the *ez* package. Effect sizes for ANOVAs were Generalized Eta-Squared (*ges*, e.g., Olejnik & Algina, 2003). Post-hoc analyses consisted in pairwise t-tests with Bonferroni's adjustment method (with the *rstatix* package). Eye-tracking variations and

raw sizes as well as performances were analysed using repeated measures ANOVA with the "group" factor as a between subject variable and the "phase" factor as a six level intra-individual variable. A 2 x 6 mixed analysis of variance was then conducted on the pupil size variations from stage S1 with a between-subject factor (Training outcome, High or Medium) and a within-subject factor (Stage, from S1 to S6). When Mauchly's test assumption of sphericity was violated, degrees of freedom of within subject factors were corrected using Greenhouse-Geisser estimates of sphericity and corrected *p*-values were provided. Given the non-normal distribution of raw pupil sizes, the analyses were made using Mann-Withney U test and Bonferroni corrections for *p*-values.

Results

Pilot Training Outcome

Among the twenty pilot students, all of them succeeded the practical airline pilot training but some of them received additional training (see Figure 2). After discussions with flight instructors, a threshold of three hours of additional practical training was set to contrast those students who faced no or little difficulty during training ($n = 14$, labelled "High level") and those who faced more difficulties during training ($n = 6$, labelled "Medium level"). Among the fourteen "High level" pilot students, twelve of them received no additional training and the two others received one and 1.5 hour respectively. The six "Medium level" pilot students received additional training because they were "below standard at a progress check" ($n = 1$, receiving 3 hrs additional training), because they failed the commercial pilot license at the first attempt ($n = 2$, receiving 3 and 5 hrs additional training, respectively), because of "problems visualizing their position in a three dimensional space" ($n = 1$, receiving 5 hrs additional training), "lack of mental resources during the flight" ($n = 1$, receiving 9 hrs additional training) and "difficulty in workload management" ($n = 1$, receiving 11.5 hrs additional training). However, the additional

training hours were sufficient to enable final success of the airline pilot training for all of the "Medium level" students.

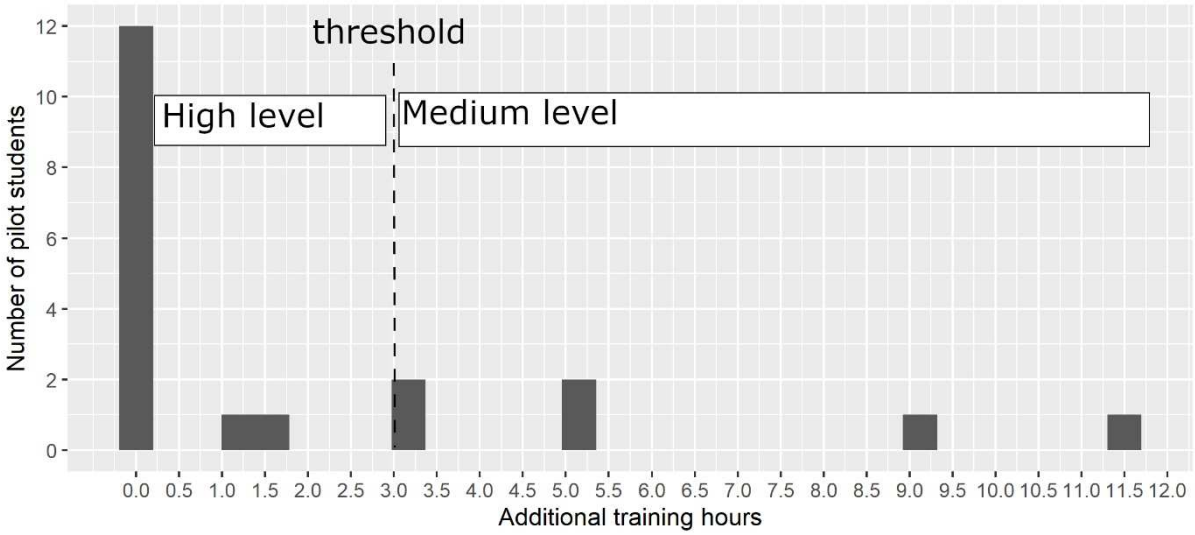

Figure 2: Barplot of additional training hours during practical training for the twenty pilot students. An arbitrary threshold of three hours of additional training has been set to identify those students who experienced some difficulties. Fourteen pilot students received less than three hours of additional training ("High level") and six of them received three hours of additional training or more ("Medium level").

Pilot Training Outcome and Task-evoked Pupil Variation

A mixed ANOVAs was run on the pupil size variation from S1 with one between-subject factor (Training outcome, with two levels, High or Medium) and one within-subject factor (Stage, with six levels, from S1 to S6). A significant interaction between the factors Training outcome and Stage was observed for the pupil size variation from stage S1, $F(3.05,54.9) = 6.10$, $p = .001$, $ges = 0.10$ (see Figure 3). The simple effect of Training outcome was significant for High level ($p < .001$) and Medium level students ($p < .001$). However, pairwise comparisons showed that the pupil size variation from stage S1 increased significantly from S1 to S4 only for High level pilot trainees (S2 vs S1, $p = .002$; S3 vs S2, $p = .038$; S4 vs S3, $p = .008$). Differences for successive stages were all non-significant for Medium level pilot trainees (the

only significant difference was between S1 and S4 and between S1 and S5). For information, conclusions were identical after changing the criterion for categorizing High vs. Medium pilot students (with 0 additional hours and with 5 additional hours). Therefore, the variation in pupil size over levels of mental workload was larger for the High level than for the Medium level pilot trainees. Thus, *H1* was confirmed: High level pilot students showed greater pupil size changes from low-load to high-load stages than medium level pilot students. Moreover, the two pilot students who experienced difficulties in the management of mental resources, as explicitly identified by their flight instructors, received 9 and 11.5 hrs additional training and would have been correctly classified in the Medium pilot student group. The difference between the two groups was the largest at stage S4, i.e. when the mental workload was at its highest level. Inspection of individual patterns revealed that the pupil size variation from S1 at S4 comprised between 0.36 and 0.67 mm for all but one High level pilot student and that all but one Medium level pilot student presented a pupil size variation from S1 at S4 of between 0.17 and 0.26 mm (see Figure 6, Appendix 1). Hence, based on a threshold at 0.30 mm for example, in each group all but one pilot student would have been correctly classified in the "High" or "Medium" category. Hence, the two pilot student groups could be quite clearly discriminated regarding each individual pupil size variation from S1.

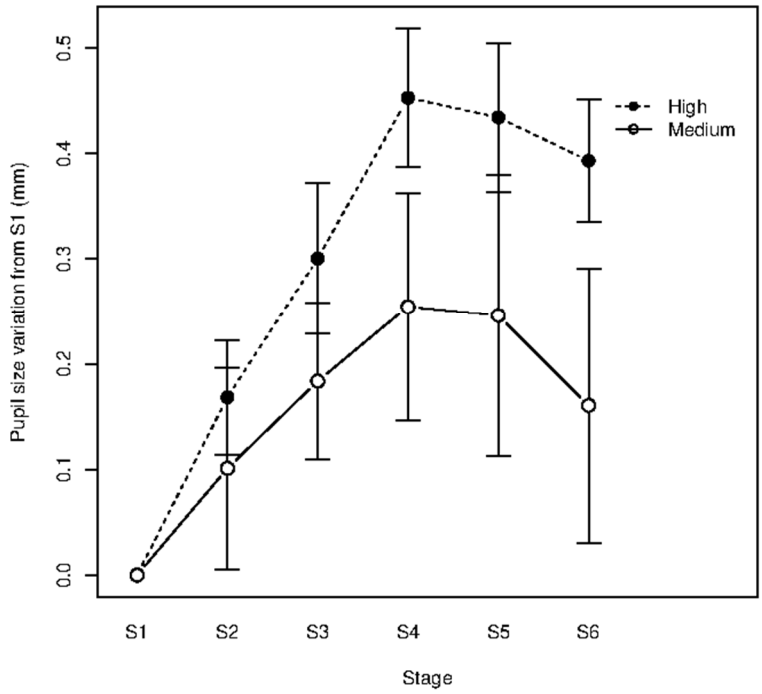

Figure 3: Pupil size variation from mean pupil size at stage S1 to mean pupil size at another stage of the Priority Management Task for pilot students who experienced difficulties during the 2-years practical training (Medium level) or not (High level). The number of subtasks increased from one to four from S1 to S4 and remained at four from S4 to S6. Error bars represent 95% confidence intervals of the means. For each subject, data were averaged in blocs of one minute.

Table 1. Means and standard deviations of raw pupil diameters in millimetres for each stage (S1 to S6) of the multitasking scenario and for each group of pilot students (Medium and High).

Stage	High pilot students *M* (*SD*)	Medium pilot students *M* (*SD*)
S1	7.52 (0.51)	8.08 (0.38)
S2	7.69 (0.51)	8.18 (0.46)
S3	7.82 (0.49)	8.27 (0.41)
S4	7.98 (0.52)	8.34 (0.43)
S5	7.96 (0.51)	8.33 (0.43)
S6	7.92 (0.50)	8.24 (0.45)

Analysis of raw pupil sizes (see Table 1 for descriptive statistics) revealed that at the lowest load stage, the High level pilot students showed lower pupil sizes than Medium pilot students (see Figure 4), $W = 145$, $p < .001$ (Mann-Whitney test). Thus, *H2* was also confirmed. For information, the five other paired comparisons for stages S2 to S6 highlighted lower pupil sizes for the High level group except for the S6 stage (Mann-Whitney tests with the Bonferroni correction for the *p*-value computation). Thus, differences were not limited to the lowest level of workload but this does not contradict our hypothesis. The global analysis of variance revealed a significant interaction between the factors Training outcome and Stage for the raw pupil size, $F(3.12,56.07) = 6.37$, $p < .001$, $ges = 0.01$, confirming the difference in pattern of pupil size variation across the task between High and Medium pilot students. Individual data (see Figure 7, Appendix 2) did not reveal the same discrimination between raw pupil diameter patterns as for the pupil size variation from stage S1 patterns. Indeed, the two groups of pilot students were less clearly discriminated with the raw pupil sizes than with the pupil size variation from S1.

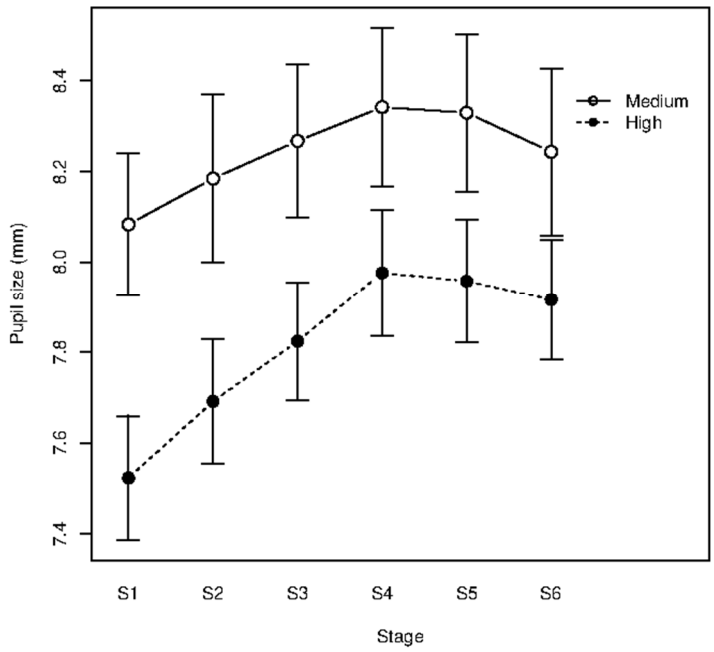

Figure 4: Raw mean pupil size (mm) at each stage of the Priority Management Task for pilot students who experienced difficulties during the 2-yrs practical training (Medium level) or not (High level). Error bars represent 95% confidence intervals of the means. For each subject, data were averaged in blocs of one minute.

Pilot Training Outcome and Multitasking Performance

A question that arose is whether these differences in pupil dilation for the two groups of pilot students would have also been observed for performance measurements. No significant interaction between the factors Training outcome and Stage was observed, $F(2.10,37.79) = 1.76$, $p = .18$, $ges = 0.05$. Moreover, only a marginally significant difference of multitasking performance was found regarding Training outcome, $F(1,18) = 3.10$, $p = .095$, $ges = 0.06$ (see Figure 5 and Table 2 for descriptive statistics). Besides, the Stage factor was significant, $F(2.10,37.79) = 115.61$, $p < .001$, $ges = 0.79$. The slight performance difference at stage S5 was confirmed with additional data and published elsewhere (Matton & André, 2014)

and confirmed by the individual performance data (see Figure 8, Appendix 3). Interestingly, performance differences between the two groups of pilot students were small (except at stage S5), compared to differences in pupil diameter patterns. Moreover, as the maximum performance score is 100 at each stage, performance data is consistent with the claim that stage S1 is an easy stage.

Table 2. Means and standard deviations of performance score for each stage (S1 to S6) of the multitasking scenario and for each group of pilot students (Medium and High).

Stage	High pilot students *M* (*SD*)	Medium pilot students *M* (*SD*)
S1	97.8 (0.5)	97.3 (1.1)
S2	97.3 (0.9)	96.5 (1.7)
S3	95.1 (1.3)	94.5 (1.8)
S4	87.6 (3.1)	86.2 (2.6)
S5	84.9 (5.0)	80.3 (5.4)
S6	88.2 (3.0)	86.7 (3.2)

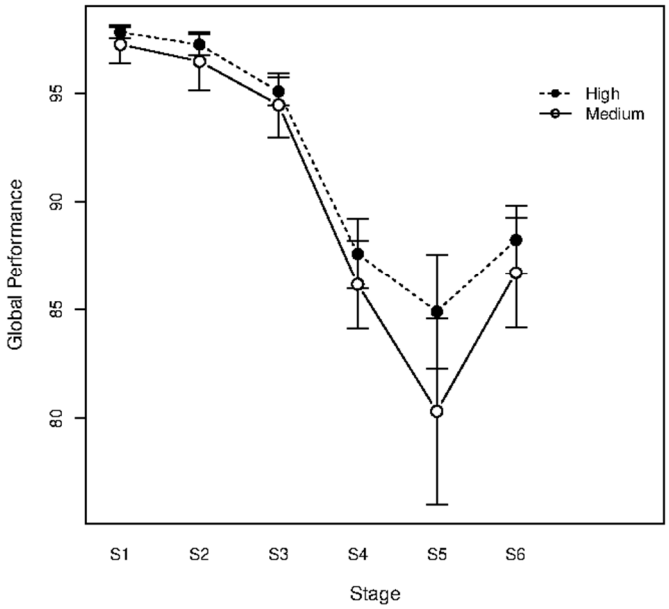

Figure 5: Global performance (average of subtask performance) for the Priority Management Task for pilot students who experienced some difficulties during the 2-yrs practical training (Medium level) or not (High level). The number of subtasks increased from one to four from S1 to S4 and remained at four for S4 to S6. Error bars represent 95% confidence intervals of the means.

Discussion

The present study investigated the predictive validity of task-evoked pupillary responses for air pilot student selection. The task used for the pupillary recording was a multitasking laboratory scenario where the mental workload increased with the number of subtasks to perform concurrently (from one to four), the Priority Management Task. Pupil size changes were computed at each stage by subtracting the average pupil diameter at the lowest-load stage (single subtask). At the end of the flight training, a group of proficient pilot students (i.e., "High level") was identified and consisted of those who needed no or only few additional flight training hours. The other group of less proficient pilot students (i.e., "Medium level") contained the students who needed at least three hours of additional flight training. Proficient pilot

students should be able to use their mental resources efficiently during real flights. Results highlighted that this group of proficient pilot students displayed a different pattern of pupillary responses compared to the less proficient pilot students. More precisely, when the task demand increased, the pupil diameters of the proficient pilot students increased more steeply compared to those of the less proficient pilot students (*H1*). Moreover, average pupil diameters at the low-load stage were smaller for the more proficient pilot students (*H2*). Thus, both hypotheses were validated.

The results are consistent with a cognitive resources perspective (see Wickens, 2002; 2008). Indeed, the least and most proficient pilot students had similar levels of performance on the Priority Management Task from low-load to high-load stages, whereas both groups differed in pupil response patterns. For a similar performance, most proficient pilot students seemed to invest fewer cognitive resources as reflected by a lower level of pupillary response. These results are also in line with EEG measurements obtained using the same Priority Management Task (Puma, Matton, Paubel, Raufaste, El-Yagoubi, 2018). Indeed, Puma et al.'s (2018) findings supported an interpretation of greater involvement of cognitive resources for this task for the lower performers compared to the higher performers. Furthermore, our results are compatible with previous findings about the relationship between individual differences in brain activation and task difficulty (e.g., Di Domenico, Rodrigo, Ayaz, Fournier & Ruocco, 2015; Doppelmayr et al., 2015; Dunst et al., 2014; Lipp et al., 2012). Our findings are also in line with the neural efficiency hypothesis (e.g., Neubauer & Fink, 2009). Indeed, this hypothesis posits that, for easy tasks, more able individuals should invest less cortical brain areas than less able individuals, which is consistent with our observations on raw pupil sizes at the lowest load stage (S1). For more complex tasks, the neural hypothesis states that more able individuals should "invest more cortical resources" (p. 1021). If we strictly consider raw pupil sizes, we did not observe higher raw pupil sizes for high performers at the highest load stage (S4). Thus,

one interpretation could be that the pilot students sufficiently practiced the task before taking the selection tests, allowing to develop efficient strategies to deal with it. Taken together with the results on the pupil size variation from the easiest stage, our findings may suggest that the most proficient pilot students were especially efficient during the easiest stage of the task. We lack data with a totally new task for all the pilot students. An alternate explanation for the raw pupil size findings is sampling error. Indeed, the Medium group may have higher trait-level raw pupil sizes than the High group. Therefore, analyses of raw pupil diameters may be difficult to interpret, and it is generally recommended to analyse pupil size variations.

The differences in pupil size change for both groups could be associated with physical limitations for the pupils of the less proficient students. Indeed, if pupil size is already large, the potential for increase is limited. However, Beatty and Lucero-Wagoner (2000, p.149) have argued that task evoked pupil responses appeared "to be independent of baseline pupillary diameter". Our results do not enable us to disentangle both interpretations (spare resources or physical limitations). Nevertheless, the group of proficient students were more able to deal with the complex situations encountered during flight training than the less proficient students. Thus, the proficient students were more likely to be able to recruit mental resources during real flights than less proficient students. Consequently, differences in task evoked pupil responses were associated with ecologically valid flight training outcomes.

Concerning individual pupillary responses (Figures 6 and 7), the question remains open as to why the patterns of the two pilot student groups were more distinguishable with the change of pupil size from the low-load stage than for the raw pupil sizes. Tsukahara et al. (2016) confirmed the influence on baseline pupil size of individual variables such as recent nicotine or caffeine consumption, or the number of hours of sleep. Thus, raw pupil sizes were potentially more affected by such variables. Pupil size change variables are less dependent on such

variables. The slope of the change in pupil size when the workload increases seems to lead to fewer false detections than the intercept.

Limits

Heitz et al. (2008) found that high-span individuals had larger resting pupil diameters than low-span individuals (for pre-experimental and pre-trial pupil baselines). Recently, this result has been replicated: individuals of the upper quartile of a working memory capacity composite score had larger baseline pupil sizes (pre-experimental and pre-trial pupil baselines) than individuals of the lower quartile (Tsukahara, Harrison & Engle, 2016). Unfortunately, we did not collect resting baseline pupil measurements. This will be the case for future investigations.

We observed the expected patterns when contrasting proficient and less proficient pilot students. However, all the pilot students of this study finally completed the flight training. The question remains open whether pilot students who fail the training would have a different pattern of pupil size variation or a similar pattern to the less proficient students.

Conclusion

The encouraging results of this longitudinal study open a new means to increase the predictive validity of measurements obtained during a selection process. Indeed, physiological data could complement behavioural performance data in order to discriminate against applicants according to their mental efficiency. For instance, physiological measurements could be collected at the end of a multistep selection process in order to maximize the likelihood of selecting applicants who will eventually succeed the training without any complementary flying hours. Moreover, physiological measurements could also be used to anticipate the need for more intensive training for specific students and/or the need for a specific training for a more efficient use of their mental resources.

Appendices

Appendix 1

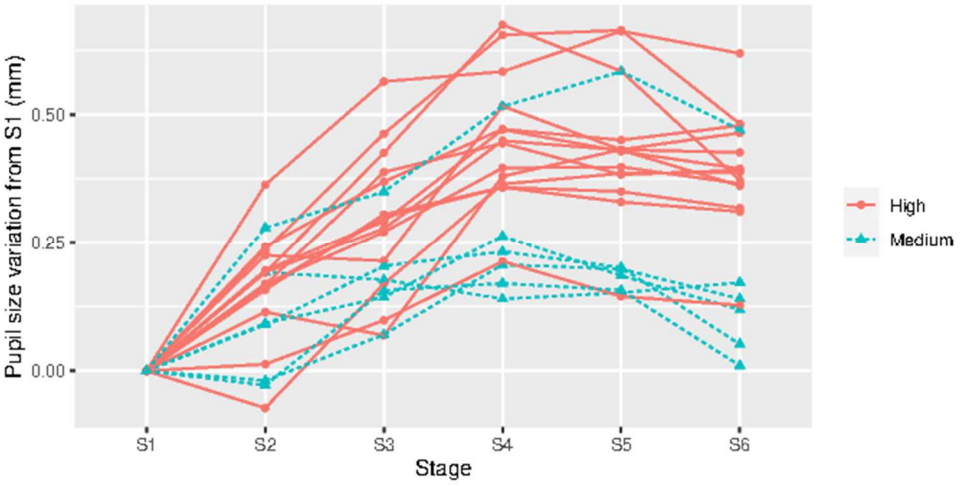

Figure 6: Individual data of pupil size variation from stage S1 of the Priority Management Task for pilot students who experienced difficulties during the 2-yrs practical training (more than 3 hours of additional training) or not (less than 3 hours of additional training).

Appendix 2

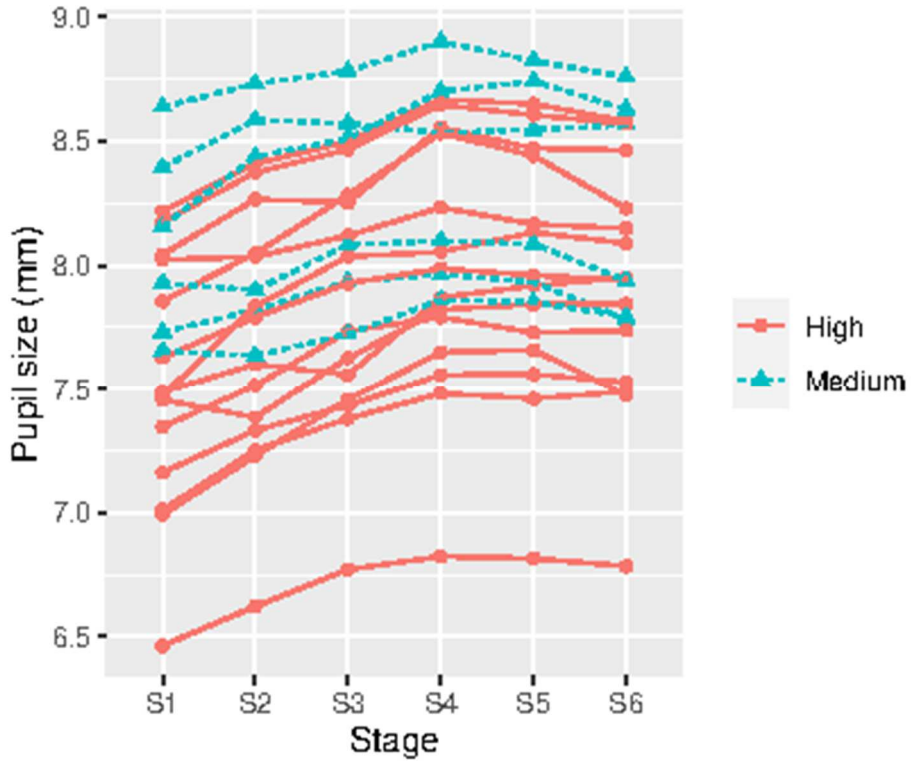

Figure 7: Individual data of raw mean pupil size at each stage of the Priority Management Task for pilot students who experienced difficulties during the 2-yrs practical training (more than 3 hours of additional training) or not (less than 3 hours of additional training).

Appendix 3

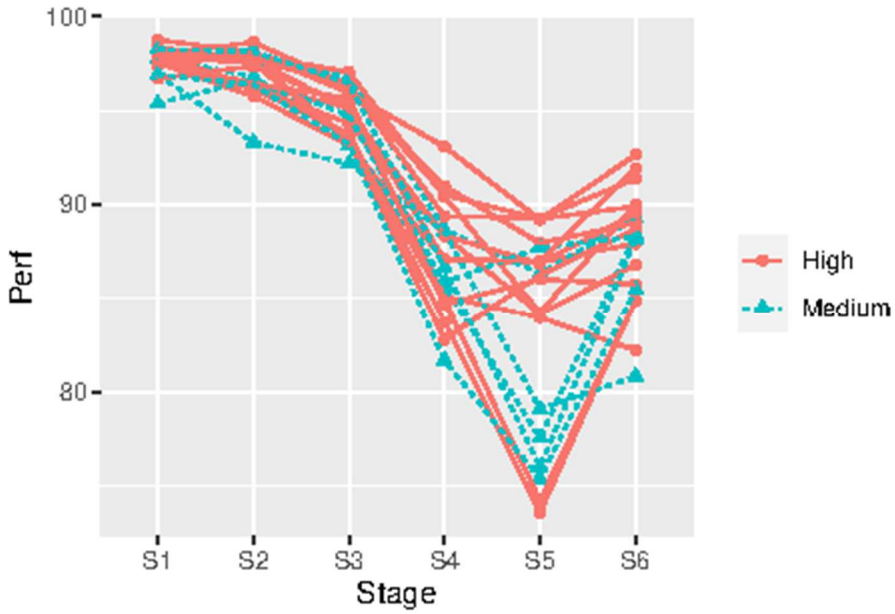

Figure 8: Individual data of performance at each stage of the Priority Management Task for pilot students who experienced difficulties during the 2-yrs practical training (more than 3 hours of additional training) or not (less than 3 hours of additional training).

Key points

- The most proficient pilot students showed a greater pupil size increase when workload increased during a laboratory multitasking scenario than less proficient pilot students.
- The most proficient pilot students had lower pupil sizes on average when workload was low, compared to the less proficient pilot students.
- Pupil size analyses might complement behavioural measurements during pilot selection in order to assess mental efficiency.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*, *205*(4412), 1289-1292.

Carretta, T. R., & Ree, M. J. (1994). Pilot-candidate selection method: sources of validity. *The International Journal of Aviation Psychology*, *4*(2), 103-117.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology*, *2*(142-162).

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45(4),* 602–607. doi:10.1111/j.1469-8986.2008.00654.x

Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine, 110(2),* 111–124. doi:10.1016/j.cmpb.2012.10.021

Damos, D. L. (1993). Using meta-analysis to compare the predictive validity of single-and multiple-task measures to flight performance. *Human Factors*, *35*(4), 615-628.

Di Domenico, S. I., Rodrigo, A. H., Ayaz, H., Fournier, M. A., & Ruocco, A. C. (2015). Decision-making conflict and the neural efficiency hypothesis of intelligence: A functional near-infrared spectroscopy investigation. *Neuroimage*, *109*, 307-317.

Doppelmayr, M., Klimesch, W., Sauseng, P., Hödlmoser, K., Stadler, W., & Hanslmayr, S. (2005). Intelligence related differences in EEG-bandpower. *Neuroscience Letters*, *381*(3), 309-313.

Dunst, B., Benedek, M., Jauk, E., Bergner, S., Koschutnig, K., Sommer, M., ... & Freudenthaler, H. (2014). Neural efficiency as a function of task demands. *Intelligence*, *42*, 22-30.

Gopher, D., Weil, M., & Bareket,T. (1994) Transfer of skill from a computer game trainer to flight. *Human Factors*, *36*, 387-405.

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology*, *45*(1), 119-129.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190-1192.

Huelmann, G., & Oubaid, V. (2004). Computer Assisted Testing (CAT) in aviation psychology. *Aviation psychology: Practice and research*, 123-134.

ICAO Circular 333. *Global Air Transport Outlook to 2030 and trends to 2040.* Available on the Internet: <http://store1.icao.int/index.php/global-air-transport-outlook-to-2030-and-trends-to-2040-cir-333-english-printed.html>

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology, 77(1)*, 1–7. doi:10.1016/j.ijpsycho.2010.03.008

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583-1585.

Kang, O. E., Huffer, K. E., & Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PLoS One*, *9*(8), e102463.

Lipp, I., Benedek, M., Fink, A., Koschutnig, K., Reishofer, G., Bergner, S., Ischebeck, A., Ebner, F. & Neubauer, A. (2012). Investigating neural efficiency in the visuo-spatial domain: an FMRI study. *PLoS One*, *7*(12), e51316.

Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. *The International Journal of Aviation Psychology*, *8*(1), 33-45.

Matton, N., & André, F. (2014). Flight training predictive validity of attention sharing with changing priorities. Oral presented paper at the *Advances International Conference on Human-Computer Interaction in Aerospace*. Santa Clara, California, USA. July 30-August 1.

Matton, N., Paubel, P., Cegarra, J., & Raufaste, E. (2016). Differences in Multitask Resource Reallocation after change in task values. *Human Factors, 58*,1128-1142.

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience & Biobehavioral Reviews*, *33*(7), 1004-1023.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies, 59(1-2),* 185–198. doi:10.1016/s1071-5819(03)00017-x

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291.

Puma, S., Matton, N., Paubel, P., Raufaste, É., El-Yagoubi, R. (2018) . Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology.* 111-120.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive psychology*, *91*, 109-123.

Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., ... & Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology*, *47*(1), 158-169.

Nadine Matton is associate professor at ENAC (Ecole Nationale de l'Aviation Civile) and is also affiliated to the CLLE laboratory. She received her PhD in cognitive psychology in 2008 at the University of Toulouse. Her research interests focus on human factors in aviation, specifically on individual differences in cognitive skills.

Pierre-Vincent Paubel is a research engineer in computer science and HF/E. He obtained his PhD in cognitive psychology in 2011 at the University of Toulouse, and he joined the CNRS as an engineer in 2016 after five years of postdoctoral positions.

Sébastien Puma is associate professor at the School of Education of Cergy-Pontoise. He was awarded his PhD in cognitive psychology in 2016 from the University of Toulouse, France. His research questions focus on models of cognitive load during learning sessions; and on ways to measure it.