# Investigating Ocular and Head-Yaw Measures as Indicators for Workload and Fatigue under Varying Taskload Conditions

Lothar Meyer, Maximilian Peukert
Air Navigation Services of Sweden (LFV)
Research & Innovation
Norrköping, Sweden
firstname.lastname@lfv.se

Tatiana Polishchuk, Christiane Schmidt
Communications and Transport
Systems, Linköping University
Norrköping, Sweden
firstname.lastname@liu.se

***Abstract*—Both workload and fatigue are decisive for human performance in current air traffic control (ATC), and, thus, should closely be monitored to ensure safety. Well-validated self-assessment and secondary-task performance measures are available but are impractical for operational monitoring because of intrusiveness and low efficiency. To overcome this gap, we investigate ocular measures and head-yaw based on eye tracking as potential non-intrusive indicators of workload and fatigue in ATC. For validation, we conduct human-in-the-loop simulations with licensed tower controllers in both single and multi remote tower working conditions. Qualitative and quantitative comparisons with conventional reference measures of workload and fatigue reveal that, among others, the eye (blink) opening speed and head yaw speed is a potential indicator of workload. Moreover, we confirm blink closing & opening amplitude as well as blink closing speed with reservations. Blink duration and blink opening amplitude may qualify as a fatigue indicator.**

***Keywords*—Workload, Fatigue, Indicator, Eye Tracking, Remote Tower**

## I. Introduction

Safety is the primary concern of air navigation service providers (ANSPs). The probability of error is increased under high- and low-workload (over- and underload) as well as fatigue conditions. Thus, we strive to ensure that air traffic controllers (ATCOs) operate within workload and fatigue thresholds—which enables us to comply with Regulation (EU) 2017/373 [1].

Measurements of workload and fatigue suffer from crucial problems: using operator self-assessment with queries is intrusive, hence, this is not feasible as a permanent instrument during operations; social-desirability bias may entice operators to understate their workload or fatigue; and the subjective measures are imprecise and not calibrated.

In this paper, we investigate several candidate indicators of workload and fatigue under varying working conditions (and with different ATCOs). By this, we aim at the management of workload and fatigue and the avoidance of safety-critical working conditions. This work contributes to an evidence-based approach that is of increasing importance with higher levels of automation or the introduction of so-called paradigm-shifting innovations.

Supported by the Swedish Transport Administration (Trafikverket) and in-kind participation of LFV within the CAPMOD project.

A current example in LFV operations are Multi Remote Towers (RTs), an operational concept that enables the provision of air traffic services (ATSs) to two or more airports from a remote location by one ATCO [2]. The possibility to operate at several airports at a time yields a major change in working conditions [3]—it is expected that this will alleviate current problems of underload and fatigue at small airports with only a few movements a day. Yet, it is hard to provide objective feedback from ATCOs to system designers and decision makers—at the same time regulators require a reliable assessment to justify the safety of remote-tower operations. The potential combination of subjective ratings with independent empirical measures would increase the capability to evaluate any suggested system design as well as to manage stress and fatigue as required by Regulation (EU) 2017/373. Fatigue Risk Management (FRMS) can increase the accuracy of risk predictions through the use of empirical indicators.

In different areas, ocular and head measures have been identified as potential indicators for workload and fatigue [4]–[6]. The validation of such indicators in ATC is not as advanced as in other fields, for example, in driver assistance. Empirical indicators of fatigue have so far not been addressed.

We assume a valid indicator to quantify signs of workload and/or fatigue evoked by the prevailing working conditions. Our investigation aims at indicators tackling two different use cases: predicting workload and/or fatigue caused by (1) a transition between different setups including modifications to working position design, procedure, task definition, working time (between-conditions) and (2) current activities and task variation described as a function of time considering different ATCOs (within-subject).

We deem single- and multi-remote-tower scenarios in Remote Tower Centers (RTCs) as a suitable case because of higher task difficulty for multiple configuration compared to single. As investigated by Friedrich [7], the increase in task diffi-

culty combined with a higher number of traffic movements is expected to modify the taskload perceived by ATCOs. For validation, we compare the empirical measures with the current standard: the subjective, intrusive measurements. Consequently, we employ a simulation study with ATCOs working in an RTC under varying taskload conditions with realistic traffic load. We choose eye-tracking and body-movement measurements as our candidate empirical indicators (for eye tracking we use a Smart Eye system, which is non-intrusive, see Subsection IV-D3). We investigate the accuracy of these measurements for predicting workload and fatigue responses to variations of taskload.

## II. Related Work

### A. Workload & Fatigue Definition

**Workload** is a subjective quality, Hart and Staveland [8] defined it as *"a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance"* (while **taskload** measures the objective demands). ICAO [9] describes **fatigue** as *"a physiological state of reduced mental or physical performance capability resulting from sleep loss, extended wakefulness, circadian phase, and/or workload [...] that can impair a person's alertness and ability to perform safety related operational duties."*

### B. Fatigue

ATCOs require sufficient levels of alertness to pay continuous attention to the air traffic situation. As a factor influencing alertness, fatigue is considered a safety hazard in air traffic as it can lead to decreased situational awareness and increases the likelihood of human error and attention lapses [10]. In addition to sleep or shift-related causes [11], taskload characteristics also influence fatigue [10]. Both high and low taskload are seen as problematic. While overload may exceed the capacity of a fatigued operator [9], underload leads to boredom and monotony [12], resulting in fatigue. While no physiological measures are integrated in daily operations today, Hu and Lodewijks indicate a good fatigue-detecting quality by camera-based ocular measures [5]. Light fatigue is associated with an increase of blink frequency per time unit [13]. The transition to severe fatigue is accompanied by an increase in blink duration. Moreover, the proportion of time during which the eyes are closed at least 80% (PERCLOS) is considered a robust fatigue measure [6].

### C. Mental Workload Measurement

Mental workload cannot be measured directly. Two major approaches can be observed: subjective studies in which self-rated ATCO workload is assessed on a scale (e.g. NASA-TLX [8], ISA [14], CHS [15]), and objective studies that aim to find observable physiological and behavioral measurements that correspond to the workload currently experienced [16].

Ocular metrics based on eye-tracking measurements have been identified as a valid indicator for mental workload in both ATM applications [17] (see Peissl et al. [16] for a survey, and applications from other areas, e.g., maritime navigation [18] or nuclear power plants [19]). For background on the pupillary system, we refer to Beatty and Lucero-Wagoner [20]; in his PhD thesis, Klingner [21] provided detailed information on methods for measuring cognitive load using pupillary dilations; Yamanaka and Kawakami [22] evaluated pupil diameter as a method to evaluate mental stress by comparing it to electrocardiogram measurements; similarly, Causse et al. [23] evaluated eye-fixation measures.

Electrophysiological measurements (EEG/ERP) showed that increased mental workload was accompanied by lower P3b while instructions were given to the pilots [24]. The processing of the instructions leads to a depletion of the cognitive resources. The EEG-based index proposed in [25] is widely used for cognitive control-behavior assessment in air traffic control and other areas.

Studies [26] performed on a driving simulator uncover that head-movement parameters could be used as indicators for a quantitative evaluation of mental workload.

### D. Remote Tower Research on Workload and Fatigue

Friedrich et al. [7] used eye-tracking data from different multiple RT experiments with varied traffic load and workplace design. They aimed to analyze the influence of multiple airport configuration on information gathering, highlighting the lack of studies correlating subjective and objective measurements in a multiple RT environment. In the simulation runs, all ATCOs observed three airports: one main airport with 50% of the traffic load, and two smaller airports with 25% traffic load each. They showed that traffic load increases the subjective ATCO workload (assessed using the ISA scale), and studied whether the workload is correlated with the eye-tracking metrics they had chosen (dwell time, fixation duration, and transition frequency). Kearney et al. [27] performed a study with remotely qualified ATCOs that compared workload between tower operations in a conventional tower and multiple remote towers using NASA-TLX. The authors identified significant differences in ATCO's mental demand, temporal demand, effort, and frustration.

Josefsson et al. [28] presented a first subset of factors that potentially drive the complexity of the traffic situation RTC ATCOs have to handle. The authors of [29] validated indicators of workload predictability in conventional and remote towers (in single and multiple conditions) and showed that—while simply using the number of ATCO tasks (like clearances and communication) does not yield a necessary condition for an increase in workload rating—indicators that integrate the communication time related to the ATCO tasks are necessary conditions, that is, each increase in workload rating is accompanied by an increase in these indicators. A part of that study is a field study in a conventional tower at Bromma airport, where the average workload rating was higher in the first three hours, during which snow sweeping occurred, than in the final hour with peak traffic (with 4, 5, and 9 movements in the first three hours, and 27 movements in the final hour). This is in contrast to the results

TABLE I. Hypothesis on Candidate Indicators

| Candidate Indicator | Workload | Fatigue | References |
|---|---|---|---|
| Blink Duration | - | + | [4], [13] |
| Blink Frequency | + | + | [6], [13] |
| PERCLOS | ○ | + | [4], [6], [31] |
| Fixation Duration | - | ○ | [7], [32], [33] |
| Head Yaw Speed | + | ○ | [26] |
| Eye Blink Opening Amplitude | ○ | ○ | |
| Eye Blink Opening Speed | ○ | ○ | |
| Eye Blink Closing Amplitude | ○ | - | [4] |
| Eye Blink Closing Speed | ○ | - | [4] |
| Pupil Diameter | + | + | [34], [35] |

Note: "+" indicates a positive,"-" a negative and "○" no results found

of Friedrich et al. [7], who showed that workload increases with the amount of traffic in a multiple remote configuration.

Furthermore, to our best knowledge, fatigue has not yet been evaluated in single and multiple RT conditions.

## III. Research Hypothesis

In Sections I and II, we underlined that several empirical indicators might be promising in predicting workload and/or fatigue. We use workload as defined in [8], see Section II. The terms sleepiness and drowsiness are covered by the term fatigue[1]. Based on the literature review, we compiled a list of potential indicators that include measures of ocular and head movements and are referred to as candidate indicators. The term "ocular measures" covers those of human physiology and behavior of the eyeball (including, e.g., pupils), eye gaze and eyelids movement. Indicators related to the movements of the eyelids are referred to by the term "blink". The chosen candidate indicators and the expected behavior in relation to workload and fatigue are listed in Table I. The table entries "+" (positive) and "-" (negative) designate the type of relation with either mental workload or fatigue. The list of candidates includes known indicators, but also new potential candidates that underscore our exploratory motivation in this study. It is important to consider a possible non-linearity of indicators, which is not addressed here. This concerns, for example, high workload which the ATCO might tackle successfully (yielding longer fixations), and a high workload that the ATCO struggles to engage with because he/she is too stressed (e.g., yielding shorter fixations) [30].

## IV. Methods

### A. Evaluation Method

With the candidate indicators identified in Table I, we aim to study the validity of these indicators for workload and fatigue using human-in-the-loop simulation. The principle approach is to use reference indicators for comparison and hence validation. Our measure of workload and fatigue refer to the current standard: ISA, KSS, and PVT (see section IV-D). Our analysis methods rely on statistic correlation analysis for quantifying the coincidence of both candidate and reference indicators for workload and fatigue. Moreover, we identify concurrent increase and decrease of workload by applying sufficient and necessary conditions as performed in [29]. These conditions help identifying non-linear relations between reference and candidate indicators such as time delay. It is consistent with the philosophy that workload is a composite construct in which the ATCO's response manifests itself in one or more physiological or behavioral signs.

The exposure of professional ATCOs to both single and multiple remote-tower working conditions creates task difficulties of varying degrees and, in combination with an increased number of traffic movements at multiple airports, varying taskload (Subsection IV-C) and, in consequence, workload [7], [28]. These conditions help us to validate the candidate indicators considering different work conditions and degrees of task difficulty as part of our use case between-conditions. The second use case within-subject aims at the prediction of workload and fatigue as a function of time considering the validity of candidate indicators across several ATCOs. Between-conditions are suitable for the comparison of entire conditions over the long term, whereas within-subject addresses the temporal dynamics of variability between the points of sampling. Hence, the latter provides more details about the present situation including tasks and activities that might have evoked workload or fatigue. This is consistent with the case of system designers who wish to evaluate a change in the concept of operations and, therefore, seek to trace local peaks in workload and fatigue to their causes in the ATCO's current activities. Distinguishing and comparing these two use cases aims at considering responses of the ATCO's physiology and behavior to changing working conditions that become visible over the long run versus those that become visible within minutes.

Besides the justification for these two cases just given, there is also a statistical reason for distinguishing: The Simpson's Paradox[2] [36]. Both use cases help us to separate the two main sources of variability: experimental condition and subject. Thus, grouping data helps to avoid the paradox.

### B. Simulator-Based Data Collection

We used the SAAB simulator with the implementation of the LFV Multi Remote Tower concept of operations of 2019 that provides the operation of two airports at a time by one ATCO. The SAAB simulator consists of two working positions, a visual presentation of the out-of-the-window view, two radar screens, a digital flight strip system, a voice communication system, a visual control unit that provides keys for adjusting the visual presentation, and a multi-purpose display (used for triggering the test procedures such as secondary task testing). The visual presentation provides a display of (one to) two airports, where

[1]For the sake of simplicity, the terms sleepiness and drowsiness are combined in this study under the generic term fatigue.

[2]"Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations" [36]

the vertical field of view and the zoom factor can be adapted according to the operators needs in the present situation. This simulation provides us with a controlled environment, where we can trigger specific taskload conditions more accurately than in a field study. We ensure realistic radio communication by using two pseudo pilots. See Figure 1 for the simulation environment in multiple configuration.

### C. Independent Variable

We vary the taskload by means of three operational concepts as principle experimental conditions, exposing participants to varying taskload ranging between low and medium. We simulate two Swedish airports: Örnsköldsvik (ESNO) and Sundsvall-Timrå (ESNN). We distinguish three experimental conditions:

- **Single.** ESNO and ESNN in single configuration with control zone (CTR) and terminal maneuvering area (TMA)
- **Multi CTR (MCTR).** ESNO and ESNN in multiple configuration with CTR only
- **Multi TMA (MTMA).** ESNO and ESNN in multi configuration with CTR and TMA (this includes a bigger area of responsibility than MCTR with more clearances and, hence, more tasks)

The traffic load increases from single to multi conditions—we combine the traffic from ESNO and ESNN in multi. The number of commercial movements is varied (2 or 4 in single mode, 6 in multi mode), all scenarios feature one VFR movement and one runway inspection each. The taskload increases from MCTR to MTMA because of the additional TMA area of responsibility.

Six scenarios were developed whereby each experimental condition is represented by two scenarios with varying aircraft types, callsigns and small variances in the timing of departures and arrivals. Each simulation run lasted 65 to 70 minutes, depending on the individual control behavior of the ATCOs. The study design complies to a cross-over design with each participant accomplishing all six scenarios for a within-subject comparison of the experimental conditions.

### D. Dependent Variables

*1) Reference Workload Indicator:*

- **Instantaneous Self Assessment (ISA).** The ISA queries provide a reference variable of workload for validating the candidate indicators [14]. The ISA scale is a five-point numerical scale, which can be used to assess workload in real-time. We assessed the workload by querying the ATCOs for their verbal rating according to the ISA scale every three minutes, the self assessment was triggered by an audio signal to which the ATCOs answered verbally.

*2) Reference Fatigue Indicators:*

- **Karolinska Sleepiness Scale (KSS).** The KSS is used to measure the subjective level of fatigue at a specific point in time [37]. The KSS is a nine-point numerical scale, varying from 1 (extremely alert) to 9 (extremely sleepy—fighting sleep). The KSS is a subjective measure that we queried before (pre) and after (post) each simulation trial.
- **Psychomotor Vigilance Task (PVT).** The PVT [38], [39] is a software-based sustained reaction-time task over three minutes and participants are asked to respond to emerging stimuli on a screen as quickly as possible. Reaction time in milliseconds (ms) was logged, lapses (reaction time $>$500ms) and false starts (reaction time $<$200ms) are excluded for further analysis. Slower reaction times indicate hereby fatigue. The median reaction time per PVT run is used for further analysis. The PVT is an objective measure that we queried before (pre) and after (post) each simulation trial.

Both KSS and PVT are well-validated instruments measuring fatigue and sleepiness [40].

*3) Ocular and Head-Yaw Measures:* We used a *Smart Eye* eye-tracking system with six infrared cameras mounted in an ATCO-centered semicircle at a distance of 60 to 100 cm at the working desk (see Figure 1). The system samples the head position/movement, eyelid movements and eye-gaze data with a frequency of 60Hz providing ocular and head-yaw measures of the ATCOs, respectively: Blink Frequency, Blink Duration, PERCLOS, Pupil Diameter, Fixation Duration, Fixation Frequency, Amplitude and Speed of Eyelid Closing and Opening, Head Yaw, and Head Yaw Speed.

The eye-tracking data log was then processed using in-house developed software (implemented in Java v.15.0.2) that extracts the desired indicators from the log for time synchronization, data smoothing, visualization and further statistical analysis. For smoothing, we use a moving average with a time-linear sampling in 10-second steps.

### E. Participants

In total, we consider three licensed ATCOs with valid unit endorsements for remote towered airports Örnsköldsvik (ESNO) and Sundsvall-Timrå (ESNN), performing six simulation runs each. Hence, we have eye-tracking data for three participants with 18 runs. The participants had a mean age of 51.3 years ($SD = 7.39$), with a mean experience of 23.3 ($SD = 10.19$). For all these 18 runs we recorded ISA, PVT and KSS data.

### F. Necessary and Sufficient Conditions

We use different analysis methods, in particular, we are also interested in predicting increases and decreases of ATCO workload. As described in our previous work [29], we do not only look at correlation but at necessary and sufficient conditions:

- A measure M forms a *necessary* condition for workload increase (decrease) if every workload rating increase (decrease) is accompanied by an increase (decrease) in M.
- A measure M constitutes a *sufficient* condition for workload increase (decrease) if every increase (decrease) in M also yields an increase (decrease) in the workload rating.

If we can identify a sufficient measure M for workload increases (decreases), observing M would be enough: each increase (decrease) of M will yield—and, hence, predict—an increase

Figure 1: Simulation environment in multiple configuration.

(decrease) in workload rating. A measure that is both a sufficient condition for workload increases and decreases would yield us a perfect predictor for workload changes.

## V. Results

### A. Workload

*1) Necessary and Sufficient Conditions:* We conjecture that an increase in the workload rating is accompanied by a decrease in the Fixation Duration, we expand that by conjecturing that a decrease in workload rating is accompanied by an increase in the Fixation Duration.

While workload is measured every three minutes, we have way more data points from Fixation Duration in the eye-tracking data, even after smoothing (we smooth over the interval -180s to 0s w.r.t. the workload measurement). Hence, we consider the "trend" in the Fixation Duration of a time interval: the slope of a regression line for the Fixation Duration observations over the time interval. With that, we refine our conjecture:

1) Each decrease in workload rating (during the period from $t_i$ to $t_{i+1}$) is accompanied by a positive trend for Fixation Duration in the complete three-minute interval $[t_i, t_{i+1}]$, or in the first or last 1.5 minutes of that interval.
2) Each increase in workload rating (during the period from $t_i$ to $t_{i+1}$) is accompanied by a negative trend for Fixation Duration in $[t_i, t_{i+1}]$ or $[t_{i+1}, t_{i+2}]$ (the rationale behind considering the following period is that an ATCO anticipates later tasks and mentally prepares for them).

As an example for this hypothesis, we consider Figure 2: between minutes 27 and 30, we observe a decrease in workload. According to point 1 from our conjecture, this decrease should be accompanied by a positive trend for Fixation Duration in the complete three-minute interval $[27, 30]$, or in $[27, 28.5]$ or $[28.5, 30]$. The red regression line gives the Fixation Duration trend for this three-minute interval. We can clearly observe the positive trend.

The hypothesis holds for the multiple configuration scenarios (MTMA and MCTR), except for two of these simulation runs where we see one exception each—in these runs, we generally observe very little workload variations. Moreover, if

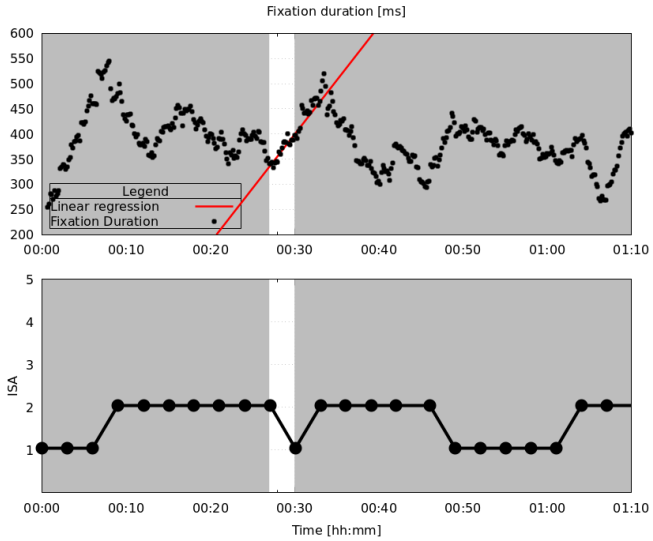


Figure 2: Example for part 1 of our conjecture: a workload decrease between minutes 27 and 30 is accompanied by a positive trend for Fixation Duration in that interval. Top: Fixation Duration data (to allow for visual distinction, we do not show all data points, but the smoothed data over the interval -180s to 0s w.r.t. the workload measurement) with a red trend for the full interval [27:00-30:00], bottom: workload measurements according to ISA scale. We highlight [27:00-30:00] with a white strip.

we have a zig-zagging pattern of the workload measurement, the succeeding three-minute period cannot be the valid criterion for workload increases (that is, workload increases must then be accompanied by a negative Fixation Duration trend in the same period), to not create a contradiction with a succeeding workload decrease.

We performed a similar analysis (necessary conditions for workload increase and decrease) using pupil diameter. However, for Pupil Diameter, we observe three exceptions (which cannot all be assigned to single or multiple configuration).

Our result suggests that the trend for the Fixation Duration can be a necessary indicator for changes in the workload rating. This would strengthen the observation from [7], [33].

However, the converse is not true: we can often observe Fixation Duration trends that are not accompanied by a workload increase/decrease, that is, we do not have a sufficient indicator. The same holds for the Pupil Diameter.

*2) Correlation Analysis:* We perform a correlation analysis according to Spearman for the quantification of the relation between ISA and the chosen candidate indicators. The indicators were samples with moving average using the same parameters as in section V-A1 (-180 to 0s). To distinguish such variances associated with the experimental conditions from those associated with the subject, we use two 2-way multilevel models for correlation. The results of the correlation analysis are shown in Table II, listing the Spearman's rank correlation coefficients ($\rho$) and p-values for the between-condition and within-subject. The multilevel correlation between-conditions and within-subjects is based on sample sizes ranging between 50 and 70 as well as between 111 and 134 respectively per group.

In the between condition, Head-Yaw Speed, Right Blink Closing Speed as well as Right Pupil Diameter show a very significant relation ($\rho = .956, p = .003$; $\rho = -.956, p = .003$; and $\rho = -.876, p = .022$; respectively). In the within-subject group, Left Blink Closing Amplitude and Left Blink Closing/Opening Speed show significant correlation with coefficients ranging between $\rho = -.1126$ and $\rho = 0.119$. Both Head-Yaw Speed, Left Blink Opening Amplitude, and Right Blink Opening Speed show a very significant relation ($\rho = .156, p = .003$; $\rho = -.150, p = .004$; and $\rho = .142, p = 0.007$; respectively).

### B. Fatigue

We perform a correlation analysis according to Spearman for the quantification of the relation between PVT, KSS and the chosen candidate indicators. The approach is similar to the ISA correlation analysis as described in section V-A2. The candidate indicators are averaged over a period of 30 min and paired with the respective KSS and PVT sample. The first 30 minutes of a run were assigned to the pre-test KSS and PVT samples and the last 30 minutes to the post-run samples. The multilevel correlation between-conditions and within-subjects is based on sample sizes ranging between 6 and 12, respectively.

The results are shown in Table II for PVT and KSS. With two exceptions, the coefficients show non-significant correlations between both reference and candidate indicators for between-condition and within-subject. The exceptions are the Blink Duration with a very significant relation to KSS for the between condition ($\rho = .938$ and $p = .006$) and the Right Blink Opening Altitude with simple significance ($\rho = .343$ and $p = .044$) for within-subject.

## VI. Discussion

In Table I, we have identified a series of candidate indicators. Our results in Section V can confirm some of them: we are able to identify promising indicators.

### A. Workload

The analysis of eye-tracking measures yields that the trend for the Fixation Duration (and Pupil Diameter) may act as a necessary indicator for workload increases and decreases, but not as a sufficient indicator: We showed that changes in workload rating, increase or decrease, are accompanied by a negative or positive trend for Fixation Duration in certain intervals, respectively. In particular, this analysis allowed us to include non-linear relations, e.g., a delayed connection (taking anticipation of tasks into account), which we do not discover by a conventional correlation test (e.g., Pearson or Spearman Correlation).

The correlation of ISA and candidate indicators yields rank correlation coefficients and significance results. While within-subject significant correlations are low (below $20\%$), between-condition correlations attain $80\%$ or more. The difference can be explained by the different number of samples that cause the correlation coefficient to appear smaller with increasing sample sizes. Nevertheless, we identified several promising indicators that behave as expected in Table I, namely Head-Yaw Speed and Right Pupil Diameter, and those where no expectation was assumed, namely Right Eye Blink Opening/Closing Speed, Left Eye Blink Closing/Opening Amplitude, and Left Eye Blink Closing/Opening Speed.

An unexpected candidate is the highly significant correlation of Head-Yaw Speed with $\rho = .956$ for between-condition and $\rho = .156$ for within-subject. An explanation for the high significance could be the additional traffic movements in multiple configuration, inducing more visual activity for information acquisition by the ATCO. Another effect originates from the Multi-Remote-Tower design, showing the airports ESNN and ESNO on the visual presentation next to each other and requiring the ATCO to yaw the head more frequently. Concluding, Head-Yaw Speed is caused by visual activity, indicating the workload evoked by visual information acquisition of the ATCO. The head-yawing is likely related to the saccadic velocity of the eye which is a known indicator of workload [16]. The group within-subject implies the correlation to even be sensitive to those changes that are associated with the condition. Moreover, the head-yaw speed is the only indicator that was independently identified as significant indicator for both ISA between-group and within-subject comparisons; this indicates a low chance of a random result. For this reason, we assume that the head-yaw speed is valid with a high probability for both subjects and conditions. Following on from this, it can be stated that both Left and Right Blink Opening Speed was identified as significant for within-subject comparisons which suggests corresponding validity.

Three (partly very) significant correlations, namely the Left Blink Opening/Closing Amplitude and the Left Blink Closing Speed, refer only to the left eye but not to the right. This asymmetry contradicts our expectation about the symmetry of ocular measurements. The most probable reason might be

TABLE II. Spearman Rank Correlation $\rho$ and $p$-values of ISA, PVT and KSS with candidate metrics ($^{*}p \leq .05; ^{**}p \leq .01$)

| | ISA | | | | PVT | | | | KSS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Candidate Indicator | between-condition | | within-subject | | between-condition | | within-subject | | between-condition | | within-subject | |
| | $\rho$ | $p$ | $\rho$ | $p$ | $\rho$ | $p$ | $\rho$ | $p$ | $\rho$ | $p$ | $\rho$ | $p$ |
| Blink Duration | .115 | .828 | −.027 | .608 | −.588 | .220 | −.019 | .916 | $\mathbf{.938^{**}}$ | .006 | .154 | .379 |
| Blink Frequency | .225 | .669 | −.085 | .103 | −.279 | .593 | −.144 | .418 | −.517 | .294 | −.271 | .116 |
| PERCLOS | .225 | .669 | −.082 | .114 | −.579 | .228 | −.120 | .501 | −.341 | .509 | −.219 | .206 |
| Fixation Duration | −.606 | .203 | −.098 | .061 | .453 | .367 | .209 | .237 | −.440 | .382 | .143 | .414 |
| Head-Yaw Speed | $\mathbf{.956^{**}}$ | .003 | $\mathbf{.156^{**}}$ | .003 | −.751 | .085 | −.246 | .169 | .086 | .871 | .076 | .670 |
| Left Blink Closing Amplitude | −.385 | .452 | $\mathbf{-.112^{*}}$ | .033 | .751 | .085 | .090 | .615 | −.237 | .651 | −.095 | .587 |
| Left Blink Opening Amplitude | −.731 | .099 | $\mathbf{-.150^{**}}$ | .004 | .127 | .810 | −.006 | .973 | .277 | .595 | −.116 | .508 |
| Left Blink Closing Speed | −.541 | .267 | $\mathbf{.106^{*}}$ | .043 | .517 | .294 | −.191 | .280 | .351 | .495 | .156 | .371 |
| Left Blink Opening Speed | .676 | .141 | $\mathbf{.119^{*}}$ | .024 | −.714 | .111 | −.237 | .178 | .362 | .481 | .066 | .705 |
| Right Blink Closing Amplitude | −.050 | .926 | −.033 | .528 | .215 | .683 | −.313 | .072 | −.362 | .481 | −.189 | .277 |
| Right Blink Opening Amplitude | −.103 | .846 | −.040 | .442 | .215 | .683 | −.121 | .495 | −.314 | .544 | $\mathbf{-.343^{*}}$ | .044 |
| Right Blink Closing Speed | $\mathbf{-.956^{**}}$ | .003 | −.070 | .180 | .751 | .085 | .163 | .359 | −.038 | .942 | .189 | .279 |
| Right Blink Opening Speed | .496 | .317 | $\mathbf{.142^{**}}$ | .007 | −.076 | .887 | .014 | .939 | −.126 | .812 | −.170 | .329 |
| Left Pupil Diameter | .595 | .213 | .052 | .333 | −.237 | .651 | −.005 | .979 | .188 | .721 | −.254 | .148 |
| Right Pupil Diameter | $\mathbf{-.876^{**}}$ | .022 | .073 | .167 | .413 | .415 | .272 | .126 | −.063 | .905 | −.065 | .717 |

asymmetries in the quality of the eye-tracking measurement that might lead to disturbing variances in the sampled indicators.

### B. Fatigue

The analysis of KSS and PVT was consistent with the correlation analysis conducted for ISA. The results yielded one candidate indicator in the between-condition case—Blink Duration—and one candidate indicator in the within-subject case—Right Blink Opening Amplitude. The increase of the blink duration with increasing KSS score rates complied perfectly with the expectations highlighted in Table I, while there was no expectation for the right blink opening amplitude. Nonetheless, the correlation is plausible and should be investigated in further detail. However, the correlations only appear in the case for the subjective KSS but not for the objective PVT. Perhaps participants did not rate their fatigue, but rather the variations in taskload. Thus, a single run was rated with a higher KSS not due to higher fatigue but due to higher monotony and a higher fatigue expectation. This biased the results, and it can be questioned if Blink Duration is rather a measure of monotony. Other potential fatigue indicators such as PERCLOS could not be proven to be related significantly by means of our data. This result contradicts [5]. A possible explanation could lie in the rather short simulation runs of the study. Since our simulations lasted only 65 to 70 minutes, this may have been too short to induce fatigue and associated physiological signs. Moreover, a measurement of participants in a laboratory setting might lead to better alertness than in real situations. As participants knew they were monitored, they were perhaps motivated to perform as a good as possible. Combined with short scenarios, this potentially resulted in such a low level of fatigue that was insufficient for validation. Additionally, the more traffic movements, the more activity contributes to avoiding underload of participants. The only condition that possibly induced fatigue was the single condition with a low number of movements, which had little effect on the results.

### C. Limitations

Generally, the quality of right-eye measurements is lower than that of left-eye measurements—given that all participating ATCOs are right-handed and the usage of the pen may visually obstruct the line of sight between right eye and the two right-hand IR cameras of the eye-tracking system. We deem this unavoidable. Our study has a couple of further limitations:

- We have a relatively low number of ATCOs. And related to this, a small number of sample points that the correlation analysis relies on. The smaller the sample size, the greater the likelihood of obtaining a spuriously large correlation coefficient in this way. For overcoming, we account only for those indicators that are significant.
- The simulation environment provides an artificial working environment, the ATCOs do not show a fully genuine working behavior.
- The environment may also yield a motivational and subjective bias on ISA, PVT and KSS.

## VII. Conclusion & Outlook

We investigate candidate indicators for workload and fatigue under varying taskload conditions stemming from single and multiple configuration in remote-tower simulations. We are able to identify a set of potential indicators—indicators that have not been investigated extensively before. To study the relation, we use both correlation tests and necessary conditions successfully. We conclude that Head-Yaw Speed, Right/Left Blink Closing/Opening Amplitude, Right/Left Blink Opening Speed, and Right Pupil Diameter are budding indicators for workload; Blink Duration is a budding indicator for fatigue; and that the trend for Fixation Duration and Pupil Diameter can be necessary indicators for changes in the workload rating. Our quantitative results are reliable because of their high maturity; the necessary-condition based results yield a direction for further validations.

In this study, we considered each of the candidate indicators separately. In future research, combining several ocular indicators may even enable us to detect subtle differences in eye movements that can be associated with changes in fatigue in

workload. In future work, we also plan to perform extensive trials including multiple configuration over longer time periods, targeting to capture increased fatigue signs and thresholds for acceptable workload.

## ACKNOWLEDGMENT

We would like to thank the participating ATCOs in RTC Sundsvall for their commitment.

## REFERENCES

[1] The European Commission. Commission Implementing Regulation(EU) 2017/373, Requirements for Providers of ATM/ANS and other ATM Network Functions and their Oversight. *Official Journal of the European Union*, 2017.

[2] A. Oehme and D. Schulz-Rueckert. Distant air traffic control for regional airports. In *IFAC HMS 2010*.

[3] M. Hagl, M. Friedrich, J. Jakobi, S. Schier-Morgenthal, and C. Stockdale. Impact of Simultaneous Movements on the Perception of Safety, Workload and Task Difficulty in a Multiple Remote Tower Environment. In *2019 IEEE Aerospace Conference*, pages 1–9, 2019.

[4] T. Åkerstedt, M. Ingre, G. Kecklund, A. Anund, D. Sandberg, M. Wahde, P. Philip, and P. Kronberg. Reaction of sleepiness indicators to partial sleep deprivation, time of day and time on task in a driving simulator – the drowsi project. *Journal of Sleep Research*, 19(2):298–309, 2010.

[5] X. Hu and G. Lodewijks. Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: The value of differentiation of sleepiness and mental fatigue. *Journal of safety research*, 72:173–187, 2020.

[6] M. L. Jackson, S. Raj, R. J. Croft, A. C. Hayley, L. A. Downey, G. A. Kennedy, and M. E. Howard. Slow eyelid closure as a measure of driver drowsiness and its relationship to performance. *Traffic Injury Prevention*, 17(3):251–257, 2016. PMID: 26065627.

[7] M. Friedrich, A. Hamann, and J. Jakobi. An eye catcher in the ATC domain: Influence of multiple remote tower operations on distribution of eye movements. In *HCII 2020*, pages 262–277, Cham. Springer International Publishing.

[8] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[9] ICAO. Doc 9966. Manual for the Oversight of Fatigue Management Approaches. 2nd Edition. Montreal, Canada, Revised in 2020, 2020.

[10] M. Bongo and R. Seva. Effect of fatigue in air traffic controllers' workload, situation awareness, and control strategy. *The International Journal of Aerospace Psychology*, pages 1–24, 2021.

[11] S. Bendak and H. S. J. Rashid. Fatigue in aviation: A systematic review of the literature. *International Journal of Industrial Ergonomics*, 76:102928, 2020.

[12] S. Straussberger and D. Schaefer. Monotony in air traffic control. *Air Traffic Control Quarterly*, 15(3):183–207, 2007.

[13] R. Schleicher, N. Galley, S. Briest, and L. Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, Jul 2008.

[14] C. S. Jordan and S. D. Brennan. An experimental report on rating scale descriptor sets for the instantaneous self-assessment (ISA) recorder. Technical report, DRA, 1992.

[15] G.E. Cooper and R.P. Harper. *The use of pilot rating in the evaluation of aircraft handling qualities*. National Aeronautics and Space Administration, 1969.

[16] S. Peißl, C. D. Wickens, and R. Baruah. Eye-tracking measures in aviation: A selective literature review. *The International Journal of Aerospace Psychology*, 28(3-4):98–112, 2018.

[17] K. K. E. Ellis. Eye tracking metrics for workload estimation in flight deck operations. Master's thesis, University of Iowa, 2009.

[18] G. Pignoni and S. Komandur. Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking. In Don Harris, editor, *Engineering Psychology and Cognitive Ergonomics - 16th International Conference, EPCE 2019*, volume 11571 of *Lecture Notes in Computer Science*, pages 106–122. Springer, 2019.

[19] Q. Gao, Y. Wang, F. Song, Z. Li, and X. Dong. Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, 56(7):1070–1085, 2013.

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

[21] J. Klingner. *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking*. PhD thesis, Stanford University, 2010.

[22] K. Yamanaka and M. Kawakami. Convenient evaluation of mental stress with pupil diameter. *International journal of occupational safety and ergonomics*, 15(4):447–450, 2009.

[23] M. Causse, F. Lancelot, J. Maillant, J. Behrend, M. Cousy, and N. Schneider. Encoding decisions and expertise in the operator's eyes: Using eye-tracking as input for system adaptation. *International Journal of Human-Computer Studies*, 125:55–65, 2019.

[24] M. Causse, E. Fabre, L. Giraudet, M. Gonzalez, and V. Peysakhovich. EEG/ERP as a Measure of Mental Workload in a Simple Piloting Task. *Procedia Manufacturing*, 3:5230–5236, 2015. AHFE 2015.

[25] G. Borghini, P. Aricò, G. Di Flumeri, G. Cartocci, A. Colosimo, S. Bonelli, A. Golfetti, J. P. Imbert, G. Granger, R. Benhacene, et al. EEG-based cognitive control behaviour assessment: an ecological study with professional air traffic controllers. *Scientific Reports*, 7(1):1–16, 2017.

[26] T. Chihara, F. Kobayashi, and J. Sakamoto. Evaluation of mental workload during automobile driving using one-class support vector machine with eye movement data. *Applied Ergonomics*, 89:103201, 2020.

[27] P. Kearney, W.-C. Li, J. Zhang, G. Braithwaite, and L. Wang. Human performance assessment of a single air traffic controller conducting multiple remote tower operations. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 30(2):114–123, 2020.

[28] B. Josefsson, J. Jakobi, A. Papenfuss, T. Polishchuk, C. Schmidt, and L. Sedov. Identification of Complexity Factors for Remote Towers. In *SESAR Innovation Days 2018*, 2018.

[29] B. Josefsson, L. Meyer, M. Peukert, T. Polishchuk, and C. Schmidt. Validation of Controller Workload Predictors at Conventional and Remote Towers. In *9th International Conference on Research in Air Transportation*, 2020.

[30] K. Holmqvist and R. Andersson. *Eye Tracking : A Comprehensive Guide to Methods, Paradigms and Measures*. Lund Eye-Tracking Research Institute, Lund, Sweden, 2017.

[31] D. F. Dinges and R. Grace. Perclos: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. *US Department of Transportation, Federal Highway Administration, FHWA-MCRT-98-006*, 1998.

[32] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung. Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, 43(1):111–121, 2001. PMID: 11474756.

[33] A. K. Faulhaber and M. Friedrich. Eye-tracking metrics as an indicator of workload in commercial single-pilot operations. In *Human Mental Workload: Models and Applications*, pages 213–225, Cham, 2019. Springer International Publishing.

[34] G. F. Wilson, J. A. Caldwell, and C. A. Russell. Performance and psychophysiological measures of fatigue effects on aviation related tasks of varying difficulty. *The International Journal of Aviation Psychology*, 17(2):219–247, 2007.

[35] P. A. LeDuc, J. L. Greig, and S. L. Dumond. Involuntary eye responses as measures of fatigue in us army apache aviators. *Aviation, space, and environmental medicine*, 76(7):C86–C91, 2005.

[36] J. Sprenger and N. Weinberger. Simpson's Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[37] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro. *Karolinska Sleepiness Scale (KSS)*, pages 209–210. Springer New York, 2012.

[38] D. F. Dinges and J. W. Powell. Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, 17(6):652–655, 1985.

[39] N. Goel, M. Basner, and D. F. Dinges. Chapter thirteen - phenotyping of neurobehavioral vulnerability to circadian phase during sleep loss. In *Circadian Rhythms and Biological Clocks, Part B*, volume 552 of *Methods in Enzymology*, pages 285–308. Academic Press, 2015.

[40] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa. Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical neurophysiology*, 117(7):1574–1581, 2006.