Validation of Controller Workload Predictors at Conventional and Remote Towers

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

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

***Abstract*—We do a field study on controller workload in a conventional tower and a Remote Tower environment (in both single and multiple mode) and give a proof of concept for the validation of indicators on their workload predictability. We analyze the number of ATCO tasks (e.g., arrivals, taxi), the communication times related to different ATCO tasks (and use them as weights for the ATCO tasks), and reaction times to SPAM queries. We show that—while the pure number of ATCO tasks is not a necessary condition for an increase in workload rating—indicators that integrate the communication time related to these ATCO tasks are, that is, each increase in workload rating is accompanied by an increase in these indicators.**

***Keywords*—Remote Towers; Workload; Human Performance**

I. Introduction

Air Navigation Service Providers aim at a well-balanced workload level for air traffic controllers (ATCOs) during all operational situations. An objective assessment of workload is crucial in order to find an appropriate level of human responsibility. However, workload is a subjective concept, which can not be measured directly. Hence, we need quantitative measures that correlate with the ATCO workload.

Remote Tower Services enable ATCOs to control traffic at various airports from a Remote Tower Center (RTC). In particular, it is possible to control several airports from a single ATCO working position (the ATCO works in "multiple mode" as opposed to "single mode").

Many studies on quantitative workload predictors exist for en-route traffic (e.g., [1]), this is not true for aerodrome control, and even less so for Remote Tower control. For staff planning, the assignment needs to ensure that no ATCO is confronted with traffic-inherent situations that constitute an unacceptable workload. While this is necessary for conventional towers, it is of even heightened interest for remote environments, because several airports may be assigned to one ATCO simultaneously. In particular, the ATCO workload in multiple mode has to take possible simultaneous events at the different airports into account. Any rostering needs information on scenarios in which extra staff is needed: when does the workload associated with the traffic of one or several aerodromes exceeds the tolerances of a moderate workload (as performance decreases for too high workload levels [2])?

Previous optimization of rosters for ATCOs in an RTC [3] used the number of Instrumental Flight Rules (IFR) flights as a measure of staff workload. However, according to LFV Operations (see [4]) only a minor part of the workload originates from IFR traffic; various other factors, like Visual Flight Rules (VFR), ground traffic movements, weather conditions, play an important role. Hence, for staff planning these factors should be integrated into a workload prediction, which is the focus of the ongoing CAPMOD project[1].

In this paper, we consider the relation between subjective workload ratings and several quantitative empiric measurements, such as the number of ATCO tasks (ATs) and measures related to the communication length. We aim to give a proof of concept for the validation of quantitative indicators on their power to predict workload in a conventional tower (Stockholm Bromma airport in Sweden) and in a Remote Tower (in a simulation environment in Sundsvall), where we consider both the control in single and multiple operation.

A. Related Work

For en-route traffic, various assessment forms of workload have been considered, see e.g. [1]. Two major approaches can be observed: subjective studies in which self-rated ATCO workload is assessed on a scale (e.g.,[5]), and objective studies that aim to find observable measures with a high correlation to an aggregation of factors that drive the complexity of an airspace (see, e.g., [6]). Pignoni and Komandur [7] recently proposed a quantitative evaluation tool of cognitive workload through eye tracking: in a field study they observed the pupil dilation for marine officers and related it to subjective workload measurements. Two significant prior studies [8], [9] attempted to assess complexity in a tower environment. Due to space restrictions we refer to [4] and Section II for additional references.

II. Methodology

In this paper, we study the relation between subjective workload ratings and several quantitative empiric measures.

[1] This research is a part of the CAPMOD (Capacity Modeling for Controller Workload Evaluation at RTC Arlanda) project supported by the Swedish Transport Administration (Trafikverket) and in-kind participation of LFV.

TABLE I: SUMMARY OF THE ADAPTED COOPER-HARPER SCALE BY DLR

Rating	Evaluation	Question for Evaluation
1	No problems, desirable	Is the situation solvable without major Disturbance?
2	Simple, desirable	
3	Adequate, desirable	
4	Small, but disruptive "delays"	Is the situation solvable by capacity-reducing measures?
5	Medium loss of capacity, which can be improved	
6	Very disruptive, but tolerable difficulties	
7	Problems to predict development of traffic situation	Is the situation solvable if the ATCO works with a reduced situational awareness?
8	Problems in information processing	
9	Problems in information reception	
10	Impossible	

For collecting the subjective workload we use the scales described in Subsection II-A. We derive quantitative measures from recorded video and communication data collected during two studies, candidate measures are, for example, the number of ATs (see Subsection II-B for a definition), the number of ATs weighted by communication-based measures (see Subsection II-C) and the response time to Situation Present Assessment Method (SPAM) queries (see Subsection II-D). In Subsection II-E we introduce how we relate the subjective workload ratings to the quantitative empiric measures. In Subsection II-F we give a description of the study setup.

A. Workload Rating

ATCOs must first of all ensure safe separation of aircraft (i.e., ensure a minimum safety distance between aircraft). In addition, they enable aircraft to reach their destinations in a timely manner. To do so, they permanently monitor air traffic, anticipate and detect (potential) conflicts and perform various other tasks that drive an ATCO's mental workload. Both taskload and workload reflect the demand of the ATCO's monitoring task: the former measures objective demands, the latter measures the subjective, mentally experienced stress during a task. Workload "represents the cost incurred by a human operator to achieve a particular level of performance"[10]. All factors external to the human operator constitute stress, which results in an individual workload, depending on different properties of each human operator [11].

For assessing workload different rating scales exist. An adapted Cooper-Harper scale (CHS) is shown in Table I, see [12]. The Instantaneous Self Assessment (ISA) scale of workload [13] uses a five-point rating scale for assessing mental workload in real time, see Table II. Because two different workload scales were used for the two observations due to different study setups, in the last column of Table II we present an approximate way of transferring one to the other.

B. ATCO Tasks

We give a number of ATs (a partly overlapping set of tasks was first defined by Massinger and Willers [14]) that describe the current traffic situation and the resulting actions of the ATCOs. The considered ATs are:

- *Arrival:* Arriving traffic that calls the tower.
- *Clearance:* Clearance for start, push back and landing.
- *Communication:* Communication occurs during clearance, weather information, with ground traffic, and during all types of non-standard phraseology, for example, questions from the flight crew.
- *Abnormal situation:* An abnormal situation indicates that a traffic situation becomes critical, e.g., infringement of separation minima. This can include that arriving traffic needs to perform a go-around because of all the traffic on the runway. An abnormal situation induces several other situations, hence, we count these.
- *Departure:* Departing traffic that calls the tower.
- *Secondary Task:* During the simulation various secondary tasks (added primary tasks, which do not affect operation, see [15]), like setting the QNH to a new value either for Örnsköldsvik or Sundsvall, were requested of the ATCOs.
- *Taxi:* Aircraft that obtained clearance for taxi.

C. Quantitative Measures Based on Communication Duration

Purely counting the ATs treats all AT types equally, some of these AT types may have a higher impact on the workload than others. One basic task—using the audio-acoustic channel—is communication, more communication results in more taskload. To reflect that we integrate the length of communication related to the AT types in the analysis. We choose the different length of communication calls for the different ATs as weights—both as average communication times for the AT types and as percentages of the total communication time.

Both values (average call duration and total communication time percentages) might indicate an increase in workload: If the individual call related to an AT takes up more time than those related to other ATs, this is caused by longer phraseology or by the increased need for callbacks to ensure proper understanding, which in turn can indicate a longer necessary time of attention for these calls. If the total time spent by the ATCOs for communication related to an AT takes up more time than that related to other ATs, we assume that the attention related to these calls increases (due to the sheer number of these calls).

D. SPAM Queries

Probe questions during experimental studies can be used to measure situational awareness. One such probe method is the SPAM [16]: ATCO reaction times to questions related to the current scenario are measured. Proper situational awareness is indicated by low latency and high accuracy.

E. Necessary and Sufficient Conditions

We aim to validate quantitative indicators on their power to predict ATCO workload, in particular, we aim to predict

TABLE II: ISA SCALE AND INTERPRETATION IN TERMS OF CHS SCALE

Rating	Workload	Spare Capacity	Description	Possible Interpretation of CHS values
1	Underutilized	Very much	Little or nothing to do. Rather boring.	1
2	Relaxed	Ample	More time than necessary to complete the tasks. Time passes slowly.	2,3
3	Comfortable	Some	The controller has enough work to keep him/her stimulated. All tasks under control.	4,5,6
4	High	Very little	Certain nonessential tasks are postponed. Could not work at this level very long. Controller is working at the limit. Time passes quickly.	7,8,9
5	Excessive	None	Some tasks and not completed. The controller is overloaded and does not feel in control.	10

increases and decreases of ATCO workload. Workload is an accumulated metric and we want to identify influencing factors (in this paper, we consider the number of ATs and ATs weighted with communication values). This leads us to look not only at correlation, but to test the predictability with criteria deviating from correlation, which still show a connection, because increases and decreases can be explained.

Borrowing mathematical notation. (see [17], [18]), we define necessary and sufficient conditions:

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

If we can identify a sufficient measure for workload increase, we can observe only the measure, and each increase will yield (and, hence) predict an increase in workload rating. Analogously, we can define necessary and sufficient conditions for workload rating decreases. A measure that is both a sufficient condition for workload increases and decreases would yield us a perfect predictor for workload changes.

F. Study Setup

We conducted studies on two different occasions at two different locations: a field study in a conventional tower at Bromma airport in Stockholm, Sweden (see Subsubsection II-F1), and an observation of a simulated Remote Tower in single or multiple mode for the airports in Örnsköldsvik and Sundsvall, the simulation took place in Sundsvall, Sweden (see Subsubsection II-F2).

1) Field Study: The study at Bromma airport was conducted on March 4, 2019 during actual operation, using five video cameras, three of which were directed toward the ATCOs (e.g., toward the flight strip board) and two were directed toward the opposite runway ends. We used the video recordings to reconstruct the ATs. During the observation two ATCOs and one assistant worked in the tower. Altogether, three ATCOs were observed for four hours. The ATCOs' mean age was 43; they had worked as ATCO for a mean duration of 19.6 years; two ATCOs were male, one female.

We assessed the workload using the adapted CHS. We queried the ATCOs for their verbal rating every 5 minutes (first every 15 minutes), resulting in a sample size of 45. Apart from the number and type of ATs (as defined in Subsection II-B) and the workload rating, we measured the length of communication calls and their purpose.

2) Simulation Study: The data collection at the Sundsvall simulation was done in weeks 19/20 2019 (May 6-17), using three video cameras directed toward the ATCOs. We used simulations of both multiple operation (of Örnsköldsvik and Sundsvall airport), and single operation of Sundsvall airport at the Remote Tower module. The observation included three ATCOs. The ATCOs' mean age was 52; they had worked as ATCO for a mean duration of 23.3 years with a mean of 5.6 years experience at the RTC; two ATCOs were female, one male. During the simulation we had five movements in singular and six movements in multiple mode. Each simulation run lasted 75 minutes. Every three minutes, we queried the ATCOs for a workload rating using the ISA scale, resulting in a sample size of 25 measurements per ATCO.

Apart from the number and type of ATs (as defined in Subsection II-B) and the workload rating by ATCOs, we measured the length of communication calls and their purpose, and the reaction time of ATCOs to SPAM queries. For measuring the reaction time to SPAM queries we took the time from the end of the query to the end of the ATCO answer, see [19]. The different SPAM queries were always introduced with the keyword "Question" and the categories were: SPAM clearance, SPAM wind speed, SPAM wind direction, SPAM QNH, SPAM altitude, SPAM position, and SPAM track.

III. Results Field Study

We analyze subjective workload ratings versus the number of ATs in Subsection III-A, and the subjective workload ratings versus various quantitative measures related to communication length in Subsection III-C.

Snow sweeping with a convoy of 10-14 vehicles appeared several times during the observation. We observed 4, 5, 9 and 27 movements during the 4 hours.

A. Workload versus Number of ATs

The number of ATs and the workload assessed by ATCOs is shown in Figure 1 in pink and ocher, respectively. We conjecture that an increase in the workload rating is always accompanied by an increase in the number of ATs in the

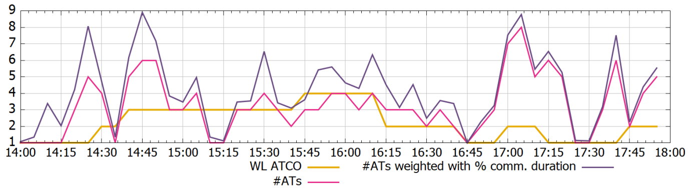


Fig. 1: Number of ATs (pink); ATs weighted with the percentage of the total communication time, see Table III (violet); and workload assessed by ATCO (ocher) for the field study at Bromma airport.

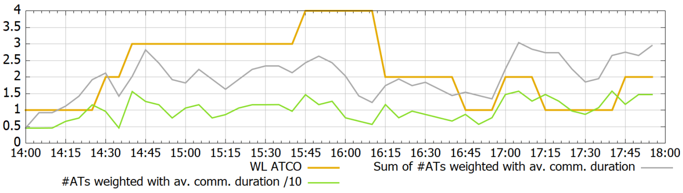


Fig. 2: ATs weighted by the average duration of radio calls related to the particular AT (green) divided by 10, the sum of this value for the current and the previous time period (gray); and the workload assessed by ATCO (ocher) for the field study at Bromma airport. See Table III for the weights.

current or previous time period (that is, an increase in rating at 14:30 is accompanied by an increase in the number of ATs at 14:25 or 14:30). The rationale behind looking at two consecutive points in time is that more ATs in one interval may accumulate and lead to an increased workload rating at the following rating query.

The conjecture holds. However, the converse is not true: not every increase in the number of ATs leads to an increased workload rating. This suggests that an increase in the number of ATs can be a necessary, but not a sufficient indicator for increased workload.

B. Communication Split: Weights for ATs

TABLE III: COMMUNICATION TIMES FIELD STUDY BROMMA

	Arrival	Clearance	Comm	Taxi
Average (in s)	10.04348	20.34783	11.2	10.7
Sum (in s)	231	468	448	321
Percentage	9.13%	18.49%	17.70%	12.68%
Range (in s)	6-16	6-57	4-72	5-28
	Departure	Ground	Total	
Average (in s)	11.44118	13.48	∅	
Sum (in s)	389	674	2531	
Percentage	15.37%	26.63%	100%	
Range (in s)	5-27	3-37		

In this subsection, we analyze the time that was spent for communication in relation to different ATs. For this the duration of each radio call and its purpose was recorded. Table III shows both the average call duration for each AT type (over all communication calls related to that AT of all ATCOs) and the sum of all radio call durations related to each AT type. Additionally, we present the latter values as percentages: All communication calls of all ATCOs during the observation accounted for 2531 seconds, out of which 231 seconds, or 9.13%, were related to arrivals.

When we consider the average duration of a single radio call for the different ATs, we can observe that clearances take notably more time than all other ATs. Each clearance is initiated by one party (usually it is issued by the ATCO), he/she obtains a reply by the other party, and for all airborne operations the second party then awaits a repetition of the information to confirm proper reception; this is not true for other call types, hence, the average duration of clearances is higher than that of other ATs. On the other hand, if we consider the absolute amount of time spent for radio calls related to the different ATs, clearances have an average value of 18.49%, while most time is taken up by communication to ground vehicles (26.63%), and very little time by radio calls related to arrivals (9.13%). In the total communication time to ground vehicles we clearly see the snow cleaning represented.

C. Workload versus Weighted Number of ATs

In this subsection, we consider both weights discussed in Subsection III-B: the percentage of the total communication time for each AT type in Subsubsection III-C1 and the average communication duration for each AT type in Subsubsection III-C2.

1) Percentage of the Total Communication Time: In Figure 1 we show the workload assessed by ATCOs (ocher), the number of ATs (pink) and additionally the ATs weighted with the percentage of the total communication time (violet) (as presented in Table III). In Subsection III-A we conjectured that an increase in the number of ATs is a necessary condition for an increase in workload. If—instead of the pure number of ATs—we consider the ATs weighted with the percentage of the total communication time, the conjecture holds (again): We can observe that an increase in workload rating is always accompanied by an increase in the weighted ATs (with percentage of the respective AT type of the total communication time) in the current or previous time period (that is, an increase of the ATCO's workload rating at 14:40 is accompanied by an increase in the number of weighted ATs at 14:35 or 14:40).

2) Average Communication Duration: In addition to using the percentage of the total communication time per AT type as weights, we use the average communication duration per AT type (as shown in Table III). Again, both the time period of workload assessment, as well as the time interval before that influence the current rating. To integrate this dependency, in Figure 2 we show—apart from the average-communication-duration weighted ATs (in green)—the sum of these values for two points in time (in gray), that is, the gray value at 15:00 equals the sum of the green values at 14:55 and 15:00.

An increase in the ATCO's workload rating is always accompanied by an increase in at least one of: the average-communication-duration weighted ATs (green) in the current or previous time period, and the sum of these for two time periods (gray). Hence, an increase in at least one of the average-communication-duration weighted ATs (green) in the current or previous time period and the sum of these for two time periods (gray) is a necessary condition for an increase in workload (at the later of the two time periods).

Still, an increase in at least one of the two criteria (the average-communication-duration weighted ATs in the current or previous time period and the sum of these for two time periods) is not a sufficient condition for an increase in workload, i.e., there exist points in time where at least one of the criteria increases, but the workload does not increase at that time period or the following time period (e.g., for the time period starting at 15:25, the average-communication-duration weighted ATs (green) increased in the previous time period, starting at 15:20, and the sum of the average-communication-duration weighted ATs for two time periods (gray) increased in the period starting at 15:25, but the workload rating did not increase in the time period starting at 15:25). Additionally, using merely the sum of the average-communication-duration weighted ATs for two consecutive time periods (gray) yields a necessary condition for an increase in ATCO workload rating (in the later time period).

Furthermore, we can observe that the sum of the average-communication-duration weighted ATs for two consecutive time intervals generally replicates the spikes and valleys in the progression of the workload rating. However, to confirm this, more observations resulting in larger data sets are needed.

D. Workload versus Weather

As a final note on the field study, we observe that the average workload rating was higher in the first three hours, during which snow sweeping occurred, than in the final hour with peak traffic (27 movements opposed to 4, 5, and 9 movements in the prior hours). More data is needed to study the influence of weather in detail.

IV. Results Simulation Study

We consider the relation between subjective workload ratings and several quantitative measures, such as the number of ATs and measures related to the communication length. As the simulation included Remote Towers both in single and multiple mode, we distinguish these categories. Again, we derive weights from the split of communication times over the different ATs (Subsection IV-A). Additionally, we consider reaction times to SPAM queries in Subsection IV-C.

A. Communication Split: Weights for ATs

We analyze the time that was spent for communication in relation to different ATs for single and multiple mode, Table IV gives the average call duration for each AT type in single and multiple mode (for each ATCO and as average over all ATCOs). Only communication shows significantly higher values in multiple than in single mode (one-sided U-test, p-value 1.65%), the other increases are not significant. The increase in average communication times related to arrivals from multiple to single was nearly significant (one-sided U-test, p-value 7.57%). Similarly, communication for clearances shows nearly significantly higher values in multiple than in single mode (one-sided U-test, p-value 6.7%). The latter is probably caused by risk compensation behavior by the operator to avoid risk at the expense of time [19], [20].

When using the average values (for single and multiple mode) as weights in Subsection IV-B, we first normalized the weights (that is, set the smallest value equal to 1, and then scaled the other values accordingly). For the observations in single and multiple mode, we use the average over all single and multiple average values, respectively. A larger measurement is needed to obtain weights that can be used in general—possibly then also for workload predictions.

B. Workload versus Number of ATs/Weighted Number of ATs

We consider the workload versus the (weighted) number of ATs, the workload was assessed every three minutes (at 9:00, 9:03, 9:06 etc.) and the number of ATs was counted from 9:00 to 9:02:59 for time point 9:00, from 9:03 to 9:05:59 for time point 9:03 etc. Thus, we shift the workload rating numbers such that they are associated with the number of ATs up to the workload assessment.

In comparison to the field study, the workload ratings for the simulation studies show smaller variation/fewer changes and long periods with the same workload assessment. This can be explained with two factors: the field study is based on the more fine-grained Cooper-Harper scale (10 values), the simulation studies are based on the ISA scale (5 values). Neither of the two studies was planned as a stress test at the boundaries of capabilities. This leaves little room for a detailed representation of the current workload using both scales, the range is, again, smaller for the ISA scale. Moreover, additional tasks like snow sweeping appeared in the field study, which led to higher variations in the taskload. The very low number of actual variations in the workload assessment hinders substantial observations.

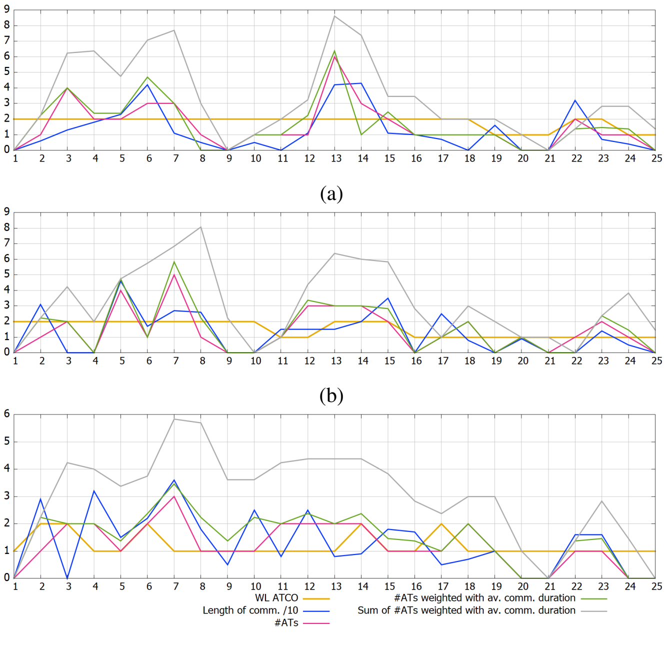
1) Single Mode: In this subsubsection, we consider the workload, the number of ATs, the length of communication during each period of observation (3 mins), plus the ATs weighted by both the average communication duration in single mode and the percentage of the total communication time of specific AT types in single mode. The progression of these values for ATCO 1, 2, and 3 is shown in Figure 3(a), (b), and (c), respectively.

All ATCOs hold an endorsement for Sundsvall, i.e., they were not confronted with a new working environment in the simulation. This explains the generally low level of the workload rating (variations between 1 and 2). The rating of ATCO 3 shows larger workload variations than that of the other two ATCOs; it can be exlained by the total time of ATCO experience: 9 years for ATCO 3 (versus 20 and 41 years).

The number of ATs is not a necessary condition for an increase in workload ($\leq$ 43% of workload rating increases were accompanied by an increase in the number of ATs). We can not observe a necessary condition for an increase in the

TABLE IV: AVERAGE COMMUNICATION TIMES SIMULATION STUDIES SUNDSVALL

	ATCO 1 single	ATCO 1 multiple	ATCO 2 single	ATCO 2 multiple	ATCO 3 single	ATCO 3 multiple	average single	average multiple
Arrival	10.83	11.5	28.5	13.67	24	9.2	21.11	11.46
Clearance	13	22.17	13.17	13.5	12.71	25.8	12.96	20.49
Comm	8.63	13.69	10.62	11.5	9.11	12.47	9.45	12.55
Taxi	12.6	8.5	8.75	5.33	20	18.2	13.78	12.04

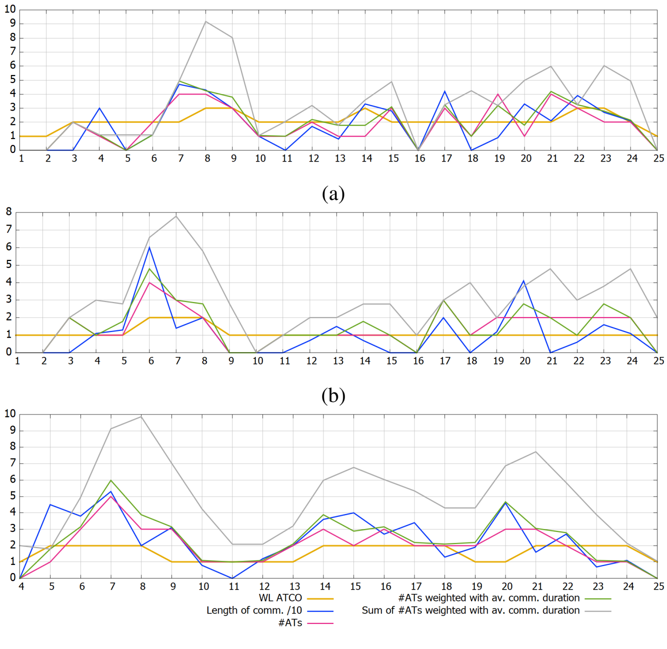


(a)

(b)

(c)

Fig. 3: Single mode: Workload (ocher); number of ATs (pink); the length of communication at each period of observation (blue) divided by 10; average-communication-duration weighted ATs in single mode (green); and the sum of the average-communication-duration weighted ATs for two consecutive time periods (gray) in single mode for (a) ATCO 1, (b) ATCO 2, and (c) ATCO 3.



(a)

(b)

(c)

Fig. 4: Multiple mode: Workload (ocher); number of ATs (pink); the length of communication at each period of observation (blue) divided by 10; average-communication-duration weighted ATs in multiple mode (green); and the sum of the average-communication-duration weighted ATs for two consecutive time periods (gray) in multiple mode for (a) ATCO 1, (b) ATCO 2, and (c) ATCO 3.

workload rating that is valid for all ATCOs. For ATCO 1 an increase in workload rating is accompanied by an increase in all measures that take the communication time into account. For ATCO 2 each increase in the workload rating is accompanied only by an increase in the sum of the average-communication-duration weighted ATs for two consecutive time periods. For ATCO 3 an increase in workload rating is accompanied by an increase in the sum of average-communication-duration weighted ATs in all but one time period. However, if we extend the condition, and do not only include the previous, but also the following period, we obtain a necessary condition: for ATCO3 each increase in workload rating is accompanied by an increase in the average-communication-duration weighted ATs in the previous, current or following time period. The rationale behind taking the following period into account is that an ATCO anticipates later tasks, and mentally prepares for them.

2) Multiple Mode: In this subsubsection, we consider the workload, the number of ATs, the length of communication during each period of observation (3 mins), plus the ATs weighted by both the average communication duration in multiple mode and the percentage of the total communication time of specific AT types in multiple mode. The progression of these values for ATCO 1, 2, and 3 is depicted in Figure 4 (a), (b), and (c), respectively. In the first two time intervals ATCO 3 was stressed due to problems with the simulation equipment, hence, we start at 9:09 instead of 9:00.

ATCO 1 has the longest experience in the RTC, but an endorsement only for Sundsvall, hence, ATCO 1 was confronted with an unknown working environment, while both ATCO 2 and 3 hold endorsements for both airports. This explains the generally higher level (and higher variations) in the workload rating of ATCO 1, who—in contrast to the other ATCOs—rated some time periods with a 3, and has a “general level” at 2, while ATCO 2 has the general level at 1, and ATCO 3 fluctuates evenly between 1 and 2.

Nevertheless, we can observe a necessary condition for an increase in the workload rating: each increase in the workload rating (for all ATCOs) is accompanied by an increase in at least one of the duration of communication at that time interval (blue) and the sum of average-communication-duration weighted ATs for two consecutive time periods (gray). This necessary condition can be compared to the necessary condition obtained for the field study in Subsection III-C2. There we identified the average-communication-duration weighted ATs in the current or previous time period and the sum of these for two consecutive time points as necessary conditions. An increase in the number of ATs is—again—not a necessary condition for an increase in workload rating. Hence, the simulation studies indicate that purely looking at the number of ATs is not enough, integrating the duration of communication yields a necessary condition. Still, in what way exactly the duration of communication constitutes a necessary condition differs slightly between the two studies, that is, either all of these need to be considered (with the formulation that an increase in at least one of these is a necessary condition for an increase in workload), or, in future research, we may aim to find a generally valid communication-length-related criterion.

For the simulation studies,—given the small data set and the human subjects—the regression results are surprisingly good; for ATCO 2 we yield an R^2-value of 0.33 and standard error of 0.27 for the number of ATs, an R^2-value of 0.51 and a standard error of 1.02 for the communication duration, an R^2-value of 0.39 and a standard error of 0.26 for the average-communication-duration weighted ATs, and an R^2-value of 0.53 and a standard error of 0.23 for the sum of the average-communication-duration weighted ATs for two consecutive time periods. This indicates that the sum of the average-communication-duration weighted ATs for two consecutive time periods can be a good predictor for ATCO workload. Also for ATCO 1 the R^2-value is high (0.47), but with a somewhat larger standard error of 0.42; for ATCO 3 we obtain only an R^2-value of 0.15 and a standard error of 0.47.

C. Reaction Time: Multiple versus Singular

Figure 5 shows the average reaction time for the three ATCOs for each SPAM query type, each in single and multiple mode. For most queries we can observe that the reaction time by an ATCO in multiple mode increases in comparison with the reaction time in single mode. In multiple mode the ATCO is confronted with more tasks, hence, he/she might be less responsive—exhibit risk compensation behavior. Due to insecurity the ATCO double checks to avoid mistakes, which results in a slowdown, which can be an indicator for uncertainty [21]. Uncertainty, apart from time pressure, is one of the main stressors. On the other hand, this trend is not true for all queries and ATCOs, e.g., the reaction time of ATCO 2 for the query SPAM track reduces for multiple against single mode, and reduces slightly for the queries SPAM clearance and wind speed; analogously, the reaction time for ATCO 3 for the queries SPAM position, wind speed and wind direction reduce in multiple mode. ATCO 1 had RTC experience, but holds an endorsement only for Sundsvall, hence, this ATCO was confronted with a new environment in multiple mode, while the other two ATCOs have endorsements for both controlled airports and RTC experience, which can explain the smaller increases from single to multiple mode, or even decreases in the reaction time for these (less time all over, less time is allocated for each task, while all tasks are fully under control). In the case of multiple RT the new working environment has the same effect as a stressor. Thus, this underlines that training in a situation helps to decrease the stress of the ATCO.

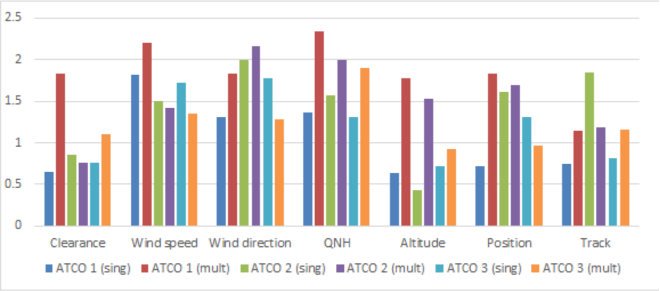


Fig. 5: Reaction times of three different ATCOs in multiple and singular mode for seven different SPAM questions (simulation studies Sundsvall).

V. Conclusion and Outlook

We studied ATCO workload in a conventional tower and in a simulated Remote Tower in single and multiple mode. As workload is an accumulated metric of different stressors, a single indicator can only partly explain the workload, while a sum of indicators can; this is particularly true for tower control, where more factors influence the stress level than for en-route control. Hence, we considered the relation between subjective workload ratings and quantitative measures that integrate more than a single indicator.

We were able to identify a necessary condition for an increase in workload rating, which holds for all increases in workload rating over all ATCO ratings for the field and simulation studies: each increase in the ATCO workload rating is accompanied by an increase in at least one of

- The ATs weighted with the percentage of the total communication time
- The average-communication-duration weighted ATs in the previous, current or following time period
- The sum of average-communication-duration weighted ATs for the previous and current time periods
- The communication duration during that time period

Thus, we validated these quantitative indicators on their predictability of workload increases. All these criteria are related to the communication time in the time period of the workload assessment (and possibly the one in the previous or following time period). In particular, we showed that simply counting the number of ATs is not a good workload indicator (it is not a necessary condition for an increase in workload rating for the simulation study), while taking the communication length into account leads to a necessary condition. We know that whenever we observe an increase in workload rating, we also have an increase in our communication-time related measure. Of course, while identifying a necessary condition gives insights into the workload development, the identification of a sufficient criterion would be even more beneficial. Our result indicates that other factors might even out variations in the communication-time related measures, e.g., the mental effort for decision-making does not yield a measurable indicator. On the other hand, the small variability in workload rating may mean that variations in our communication-time related measure in fact yield a change in workload, but the scales are not fine enough to reflect these changes. For future work, we aim also for a sufficient criterion for workload rating decreases, which would altogether yield quantitative workload predictors. This study is a proof of concept, with a relatively small data set, larger data sets for future studies are planned.

Here, we used communication data for the full study period to give weights to the different ATs, this has the advantage that the weights can be derived from large data sets and then lead (if valid), in combination with AT predictions, to predictions of workload values. Otherwise, we can take a more detailed look at the communication length related to each specific AT over time: studying correlation between the temporal progression of the communication length of an AT and the workload rating—but this could not be used for predictions.

While the regression analysis for the simulation studies points to the sum of average-communication-duration weighted ATs for two consecutive time periods as a good indicator for workload progression, the data set is too small to draw final conclusions. We aim to use these preliminary results as a base for future studies with larger data sets in different tower types.

We use the ISA scale and the CHS for workload rating. However, we can only observe relatively small variations in the workload, which impede finding correlations. Both scales work well to differentiate critical/unacceptable levels of workload from levels with non-reduced situational awareness, that is, for a binary decision. In this study, we observe that smaller variations of workload on levels with non-reduced situational awareness cannot be reflected equally well. Moreover, the ATCOs are less familiar with the scales than with the working environment. The simplicity of a numerical scale may lead the ATCOs to name a reasonable number, without mentally checking the associated verbal description of that rating. Additionally, we verbally queried for the workload rating, social desirability can cause the ATCOs to only answer with low ratings (which correspond to proper situational awareness and full control of all tasks). Hence, we see a necessity to develop an instrument that is able to register variability on lower workload levels.

Moreover, in future studies we plan to integrate measurements on other factors, e.g., runway friction values, which can be a predictor for necessary snow cleaning, which in turn—due to the high coordination effort—may impact ATCO workload.

Finally, in this paper we aim at the relation between subjective workload ratings and quantitative empiric measurements. Another approach can be to find a physical measurement (e.g., pupil diameter [7]) with a high correlation to workload, if we then obtain that the quantitative empiric measurements have a high power of predicting this physical measurement representing the workload, we may use the quantitative measurements for workload prediction.

Acknowledgment

We thank Christoffer Massinger and Hampus Willers for participating in collecting the data at Bromma airport.

References

[1] D. Delahaye and S. Puechmorel. Air traffic complexity: towards intrinsic metrics. In *Proc. of the third ATM Seminar*, 2000.

[2] R.M. Yerkes and J.D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *J. of Comp. Neurology and Psychology*, 18:459–482, 1908.

[3] B. Josefsson, T. Polishchuk, V. Polishchuk, and C. Schmidt. A step towards remote tower center deployment: Optimizing staff scheduling. *Journal of Air Transportation*, 27(3), 2019.

[4] 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.

[5] C. Möhlenbrink and A. Papenfuss. ATC-monitoring when one controller operates two airports research for remote tower centres. *Proc. Hum. Factors Ergon. Soc. Annu. Meet*, 55(1):76–80, 2011.

[6] E. Zohrevandi, V. Polishchuk, J. Lundberg, Å. Svensson, J. Johansson, and B. Josefsson. Modeling and analysis of controller's taskload in different predictability conditions. In *Proc. of 6th SID*, 2016.

[7] 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*, pages 106–122, 2019.

[8] F. Netjasov, M. Janić, and V. Tošić. Developing a generic metric of terminal airspace traffic complexity. *Transportmetr.*, 7(5):369–394, 2011.

[9] A. Koros, P. S Rocco, G. Panjwani, V. Ingurgio, and J.-F. D'Arcy. Complexity in Air Traffic Control Towers: A Field Study. Part 1. Technical report, U.S. Department of Transportation, FAA, 2003.

[10] S. G. Hart and L. E. Staveland. Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human mental workload*, pages 139–183. Elsevier, 1988.

[11] H. Bubb. Human reliability: A key to improved quality in manufacturing. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 15(4):353–368, 2005.

[12] A. Papenfuss and M. Peters. HMI Laboratory Report 8: Analysis of Critical Situations at Remote Tower Operated Airports. DLR, Institut für Flugführung, Braunschweig, 2012.

[13] 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.

[14] C. Massinger and H. Willers. En analys av den mentala arbetsbelastningen för en RATCO vid hanterbara flöden. B.S. thesis, Linköping University, 2019.

[15] Secondary tasks. https://tinyurl.com/y6hnd9lc. Accessed on 19/10/03.

[16] F. T. Durso and A. R. Dattel. SPAM: The real-time assessment of SA. In S. Banbury and S. Tremblay, editors, *A cognitive approach to situation awareness: Theory and application*, page pp. 137–154. Ashgate, 2004.

[17] Michiel Hazewinkel. *Encyclopaedia of Mathematics: Monge—Ampère Equation—Rings and Algebras*. Springer, 2013.

[18] Khan Academy. If X, then Y: Sufficiency and Necessity. https://www.khanacademy.org/test-prep/lsat/lsat-lessons/logic-toolbox-new/a/logic-toolbox--article--if-x-then-y--sufficiency-and-necessity, last accessed 16.05.2020.

[19] L. Meyer, M. Peukert, B. Josefsson, and J. Lundberg. Validation of an empiric method for safety assessment of multi remote tower. In *ATM Seminar*, 2019.

[20] G. J S Wilde. Risk homeostasis theory: an overview. *Injury Prevention*, 4(2):89–91, 1998.

[21] H. Rastegary and F. J. Landy. *The Interactions among Time Urgency, Uncertainty, and Time Pressure*, pages 217–239. Springer US, 1993.