Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019)

Validation of an Empiric Method for Safety Assessment of Multi Remote Tower

Lothar Meyer, Maximilian Peukert, Billy Josefsson
Research and Innovation
Luftfartsverket LFV
Norrköping, Sweden
lothar.meyer@lfv.se
maximilian.peukert@lfv.se
billy.josefsson@lfv.se

Jonas Lundberg
Department of Science and Technology
Linköping University
Linköping, Sweden
jonas.lundberg@liu.se

***Abstract*— The novel multi remote tower concept involves the control of two airports by one tower controller from one remote workplace at a time. In order to implement a multi remote tower into operations, a safety assessment is crucial to evaluate existing risks. Since there is currently no operational experience available concerning this concept, the hazard identification and risk mitigation remains hypothetical. However, empiric data is needed for evaluating and focusing on the safety-relevant hazards that are multi remote tower specific. To close this gap, we developed the MERASSA concept for gaining evidence on the safety-relevance of hazard using Human-In-The-Loop simulations and stress test scenarios. The method was assessed through a validation study at the multi remote tower case using eight identified hazards that are human-issue originated. In total 32 simulation runs with eight rated and experienced tower controllers were carried out. The results of the study show the ability of the tower controller to compensate risk by slowing down the work speed. No hazard could be verified through a comparison of the multi and single runway baseline scenario. Additionally, the results indicate a clear lack of confidence of the tower controller to control two airports at a time due to the need to share attention across the work environment. The comparison of the empiric and subjective results show equal trends which are a sign for the success of applying the method. However, a major drawback of using simulations and stress test scenarios are the enormous efforts needed to control the conditions of testing.**

Keywords-Safety Assessment; Socio-Technical Systems; Human Error; Multi Remote Tower; Air Traffic Control

I. INTRODUCTION

Remote tower technologies have been put into operations or are under implementation worldwide with the objective to enhance safety, capacity, and cost-efficiency of tower control services. Operational experience collected from the remote tower center (RTC) in Sundsvall is used at LFV to pave the way for developing and implementing the multi remote tower concept as a further step in regards to the mentioned key performance areas. The concept allows one tower controller to control two airports from one workplace at a time involving control, information and alerting service. To support the approval process of multi remote tower for operations, a safety assessment of the concept is an obligatory step. However, current safety assessment methodology relies on subjective statements of experts. Often these estimations are not evaluated if the identified hazards are indeed multi remote tower specific and thus safety-relevant. We developed a method to provide empiric safety evidence by verifying the safety-relevance of identified hazards pre-operationally. The method addresses primarily hazards originated in human issues that we consider as the major multi remote tower challenges. In order to understand the novel character of multi remote tower induced risk characteristics, we provide in the following an overview of the estimated differences between single and multi remote tower in regards to the impact on the tower controllers work and deficiencies from a safety and human factors perspective. The LFV concept is used as basis for our discussion.

In single runway remote tower services, LFV had the focus of designing the work position to resemble conventional towers as close as possible. Hereby, a preserving of the well-established work methods of the tower controllers was the aim. The remote tower-specific education focuses today on technical skills such as error handling and related procedures for recovery described in the "Checklista Felfunktioner" (Malfunction Checklists). The transitional education process of the tower controller from conventional services is hence less extensive since no major additional skills are needed.

In contrast, the multi remote tower concept exhibits new challenges for the tower controller. Firstly, sharing of attention and awareness across two independent operational environments is required. Thus, a need for mental switching by the controller between two airports emerges and correct allocation of traffic information could perhaps be problematic in turn [1]. Secondly, multi remote tower concepts in general experience an ongoing development that incorporates new technical functions and automation on an unprecedented scale in ATC. This concerns primarily functions to switch between the operating modes "one airport" and "two airports", so-called split-merge-functions, according to the workload perceived by the operational situation. In addition, assistance tools are implemented to schedule and harmonize the traffic flow [2] and coordination with the approach. The increasing technical

This study was funded by the Swedish Transport Administration and the LFV Air Navigation Services of Sweden through the projects MERASSA and Flygsäkerhet Fjärrstyrda Torn (FSFT)

support intensifies the symbiosis between human work and technology that the controller relies on. Taking into account the full extent of diversity of errors and the related quality of risk sources, we consider the ability to provide the safety evidence to operate multi remote tower with an Equivalent Level of Safety (ELoS) [3] as the key factor within any approval process.

Recent research addresses many of the mentioned issues through simulations and eye tracking measurements. The ability to visual monitor touchdowns and takeoffs in multi remote tower was investigated already [4]. Here, a significantly higher rate of monitoring of safety-relevant events in single remote tower in comparison to multi remote tower was detected. A large scale demonstration on multi remote tower has been done at the example of Shannon and Cork airport by means of 50 live trials and three tower controllers [5]. In result, no safety occurrences were observable. However, tendencies of incorrect button selection, difficulties to see small aircraft and increased workload as a result of simultaneous tasks were found. A well-balanced workload seems to be the primary factor of operating multiple airports. In this scope, appropriate factors were identified expressing the situational complexity as the first cause of workload [6, 7]. According to that, certain pair and triple combinations of operational events might raise criticality if occurring simultaneously. Thus, safety concerns could arise due to conflict situations that need to be handled in parallel to regular movements on the other airport. The SESAR project 6.9.3 revealed 29 possible human performance issues and 20 possible technical failures that could become safety-relevant [8]. Moreover, nine of them address information confusion errors and emphasize the critical role of the tower controller to keep track of all operational relevant information in situations where traffic situations and related conditions change rapidly on both airports in parallel. The related risk of information confusion was investigated in [9]. The study confirmed the general safety relevance of this type of error for operating multi remote tower. As known from flight deck safety research [10], it should be moreover highlighted that human errors likely follow technical failure events. This is an additional side effect of the increased human-technology interaction that multi remote tower attributes.

The safety assessment work done so far involves an analysis of risks using the Eurocontrol Safety Reference Material [11]. It bases on pre-existing hazards of the Eurocontrol Accident Incident Model [12] such as “Adverse Weather Conditions” or “Snow/slush on the runway”. The level of analysis remains generic due to the pre-industrial development phase that corresponds to maturity level v3 according to the E-OCVM model. The hypothetical level of discussion and argumentation of possible risks makes it hard to evaluate whether the hazards identified are multi remote specific and of substantial safety-relevance. This is reasonable since expert statements based on simulation trials [13] with little conclusion about the safety-relevance of hazards. However, the final implementation for putting multi remote tower to operation demands for an evidence-based commitment of the safety assessment process concerning the safety-relevance of the hazards found so far. Reliable statements are a prerequisite for evaluating the training program, education, and design of the proposed solution for implementation as part of the safety case [14]. The question remains for us if hazards that were identified, e.g. in the scope of a Functional Hazard Assessment (FHA), are significantly promoted by human performance in a multi remote tower environment compared to a single (runway) reference scenario. A change in the promotion, expressed as a frequency of occurrence, is a crucial criterion for considering any identified human performance related hazard as safety-relevant. Only an empiric approach to safety assessment is able to obtain evidence concerning the safety-relevance. This can be feedback to the FHA and used to evaluate the hazards found so far. Such an evaluation circle is known from safety management systems and the key to lifting the awareness of safety workshops participants for multi remote tower related risks to a superior level.

Following the task definition of the Eurocontrol SAM SSA[1] (System Safety Assessment) [15], the safety objective of a certain hazard is verified in the case that the tower controllers demonstrate the capability

- to compensate or to avoid the hazards that are human-error originated or
- to detect and handle technical failures

to an equal degree as compared to a single reference scenario. By such a comparison process, the pre-implementation baseline performance can be verified against post-implementation performance, to see if an ELoS-performance has been maintained [3]. As such, we consider the risk as acceptable due to the proof of operating multi remote tower under an ELoS.

The dedicated concept developed and applied is called Methods for the Empiric Risk Assessment of Socio-Technical Systems in ATM (MERASSA) [16]. It addresses the mentioned question by providing a quantified safety benchmark as safety evidence for a novel system complying with an ELoS. The methods base technically on stress test scenarios in Human-In-the-Loop Simulations that test the tendency of multi remote tower to promote hazards by means of measuring safety performance indicators. The concept was first prototyped and proofed in [9] and further developed by means of a prestudy [17].

This paper presents the results of the MERASSA validation study at the LFV multi remote tower concept. The results are produced in the scope of the second iteration evaluation that base on the training and education of the test persons gained under the first iteration of MERASSA (pre-study). The paper will, for this reason, focus on findings concerning the safety-relevance of multi remote tower-hazards. At the same time, the method will be tested at the multi remote tower case. The discussion and conclusion will follow this distinction between both aspects in which the success of applying MERASSA will be evaluated by comparing the empiric and subjective results of the simulation.

The paper will introduce the details of the concept with a focus on the work principles of test procedures and appropriate safety metrics. Thereafter, results of the safety workshop in the

[1] The objective of performing a SSA is to demonstrate that the system as implemented achieves an acceptable (or at least a tolerable) risk and consequently satisfies its Safety Objectives specified in the FHA and the system elements meet their Safety Requirements specified in the

shown. The experimental setup provides an overview on the simulation platform and the stress test scenarios with embedded test procedures. The sampled safety metrics are presented and discussed by means of the distributions and appropriate diagrams. Finally, conclusions about the multi remote tower hazards are drawn on the empiric as well as the subjective data collected.

II. THE MERASSA CONCEPT

The MERASSA-concept has the objective to provide safety evidence by verifying the hazard's safety-relevance empirically. This approach complies to the safety activities related to the SAM SSA [18, 15] as part of the pre-operational system implementation and integration. The verification is also part of an iterative process (Figure 1) that adopts the cyclic "safety risk management process" for proving the ELoS [3] as following:

1. *Identifying Hazards*: The iterative process initiates at the FHA hazard identification and assessment. Here, we identify a set of multi remote tower related hazards with the related severity estimations of the consequences and safety objectives by means of expert statements.
2. *Defining Safety Performance Indicators:* The hazards are used to define safety performance indicators[2] (SPI) that are appropriate to monitor the safety performance with regard to the hazard's safety objective. The pre-implementation baseline performance can thus be verified against post-implementation performance. Technically, these are events or state measures that are causally related to the hazard occurrence and thus indicate the related capability of the tower controller to avoid or mitigate the specific hazard.
3. *Developing Test Procedures*: For monitoring and sampling the safety performance indicators, test procedures and related safety metrics are designed for Human-In-The-Loop Simulations. A test procedure describes an SPI-specific stress test that samples the safety metrics needed for quantifying the safety performance of the operator under controlled conditions.
4. *Determining the Safety Benchmark*: The comparison of collected safety metrics shall finally allow for concluding on a systems tendency to promote a specific hazard. The verification of the safety-relevance decides if the ELoS can be assumed.
5. *Apply Risk Mitigation*: Deficiencies are mitigated using a risk mitigation workshop that feedback the corresponding measures into the system definition. This step is not presented in the scope of this paper.

For sampling the SPI, test procedures were developed following three different work principles that are explainable by means of the bow-tie diagrams (Figure 2) and are described as follows:

- *Conflict Induction Test*: This type of test consists of a conflict situation that is incorporated into an otherwise operationally normal situation. Our quality indicators of the capability to act are the time to detect the conflict as well as the compliance of the chosen solution with the operational procedures.
- *Query Test*: The query addresses any available state variables that are causally linked to the hazard of interest, including situational awareness and workload queries.
- *Secondary Task Test*: This type of test instructs the test person to accomplish a pre-specified task. The test might address the capability to handle information and equipment that are of operational relevance. Secondary tasks shall have a minimum impact on the primary task performance and take place at the work position with no interrelation to operations.

All test assumes the control of all other variables in regards to the actual operational conditions under which the test is applied. This concerns especially the timing of execution and the actual traffic situation.

A. Safety Metrics

Safety metrics are dependent variables that quantify the related Safety Performance Indicator and represent the human performance of the human in regards to safety. For defining the safety metrics, we use the definition by Bubb that defined human performance as the provision of work quality per time [19]. In this regards, we consider the Speed-Accuracy-Tradeoff (SAT) as an appropriate quantifiable behavior phenomenon describing the uncertainty of human work as an interrelation between work speed and accuracy [20]. A common explanation for this phenomenon is provided by the sequential sampling model that refers to the efforts of accumulating (sampling) information for more accurate decisions at the expense of time [21]. Since time is limited, the decision maker is forced to balance between quality and time costs of sampling information from the environment until a threshold of evidence is reached. In the ATC context, controllers use variable safety margins, such as delays and spacings, to decrease the workload in a well-balanced efficiency-safety-tradeoff [22, 23]. We consider the accuracy and related human error probability as well as its interconnection to work speed, measured as reaction time, as valid indicators of human uncertainty and work performance in ATC.

As performane measure, reaction time is distinguished into the Time-To-Detect (TTD) and the Time-To-Solve (TTS). The TTD is relevant in the conflict induction test for quantifying the test person's ability to search and identify potentially threatening situations. The TTS is used in tests in which the time to solve a task indicates the ability to decide and follow the correct procedures in the respective situation. Complementary, the human error indicates compliance of the response to the defined task.

B. Safety Benchmark

The safety benchmark provides a standard of comparison indicating differences in the safety metric between the multi remote tower (multi-mode) and the baseline single runway remote tower (single-mode). Technically, samples of safety metrics collected from experiments shall finally allow for testing statistical significance.

[2] A data-based parameter used for monitoring and assessing safety performance [33]

C. Verification Criteria for the ELoS

The criteria for considering multi-mode as equally safe as compared to single-mode is based on the benchmarks of work speed and accuracy measures. In order to account for the interrelation between both metrics, we refer to the risk compensation theory. According to that, the addition or removal of safety factors will result in a behavioral change that leads to compensatory measures involving adjustments to work speed and accuracy [24]. Risk will remain constant even if work conditions become more unfamiliar and difficult by adjusting the work speed correspondingly. Decision makers often switch strategies to reduce cognitive effort, increase accuracy, or respond to time pressure [25]. Assuming the tower controller to be familiar with work in single-mode, we consider the conditions of work in an unfamiliar multi-mode instead as an increase of difficulty. Thus, the adjustment of work speed is an indicator of successful risk compensation by the test person.

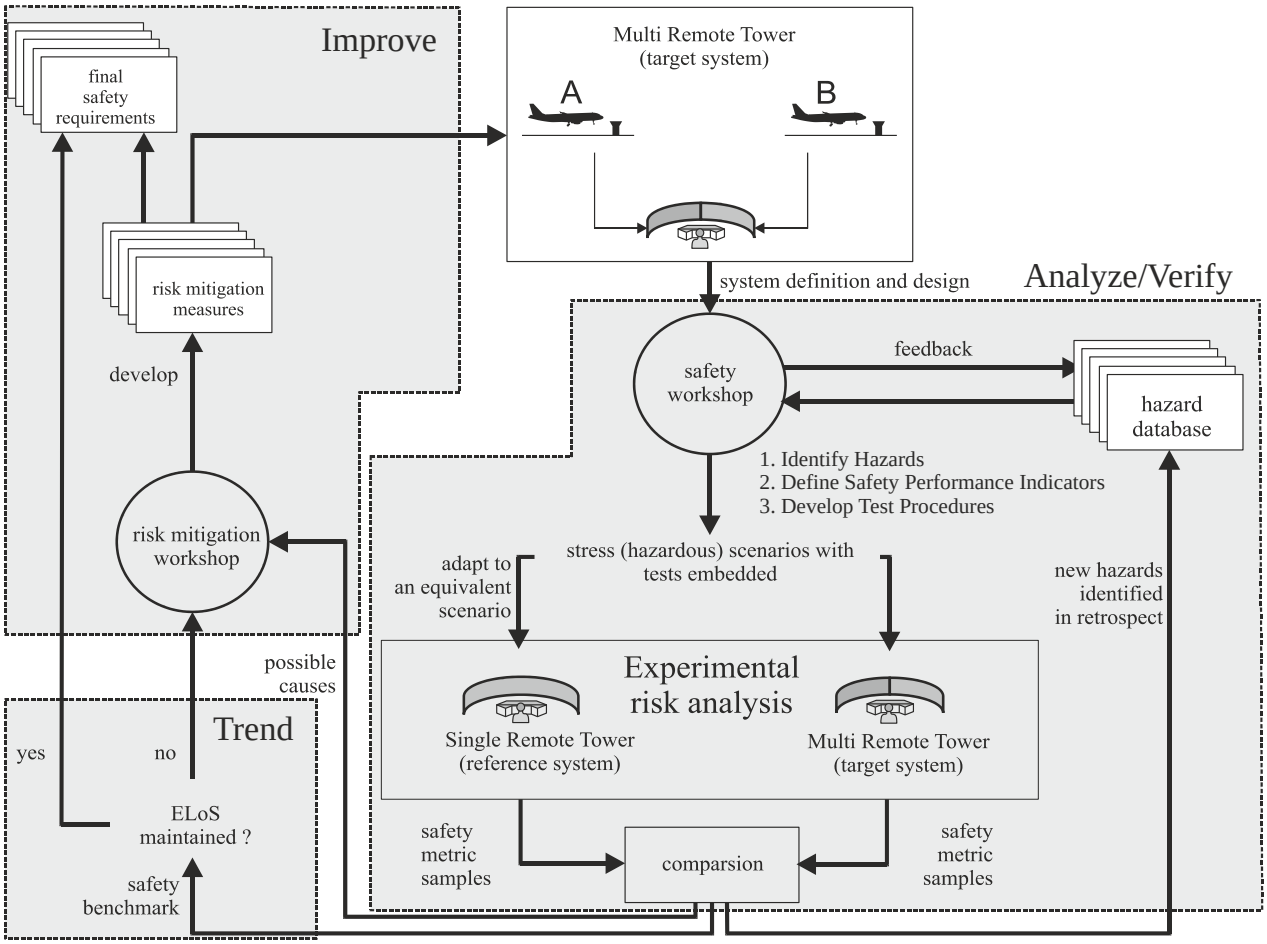

Figure 1. MERASSA Concept [16]

Based on that, we define the following cases as criteria of ELoS:

- No significant difference for reaction time and number of human error is an indication for the test person's capability to act with equal work speed and risk in both the multi- and single-mode.
- A difference of reaction times (with an equal number of human errors) indicates the response of the test persons to increased difficulty and uncertainty in multi-mode. From both directions, we consider slowing down the work speed as the correct response to an increased difficulty as defined by the risk compensation theory.

From experience, human errors are rare events where we expect only a small number of samples. Therefore, the reaction times are expected to provide more samples that indicate the potential for human error occurrence indirectly by the interrelation of the SAT.

III. HAZARD IDENTIFICATION

By means of a three-day safety workshop, three tower controllers from Stockholm-Arlanda, Stockholm-Bromma, and Östersund as well as an Airbus A320-rated Pilot identified a number of multi-mode-related hazards. The workshop was conducted in accordance with [26]. The following system was assumed during the workshop:

- Visual presentation on 14 HD screens, radar, flightstrip and voice com system.
- Control of max. two airports at a time.

- Coupled tower frequencies with optional decoupling, ground frequencies on separate frequencies.
- Limited horizontal Field-of-View of 180° for each airport. Rotation of the view to the backside by using user input commands through buttons.
- No operational restriction. The controller is allowed to apply the same procedures in multi- as in single-mode.

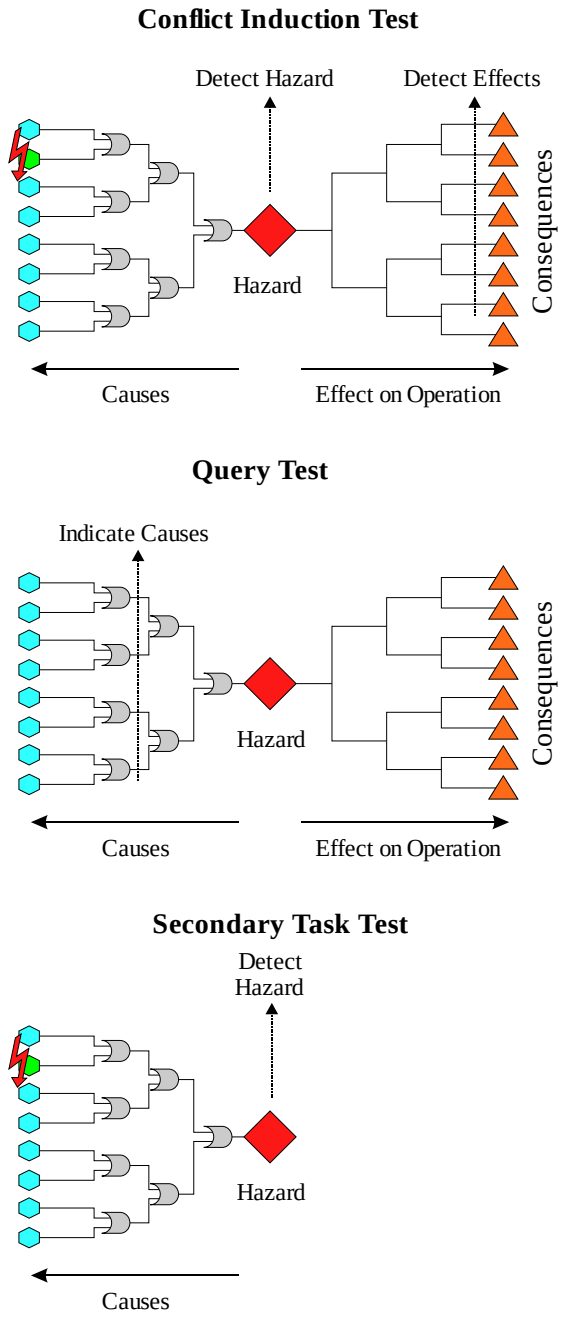

Figure 2. Test Procedures Working Principle

Table 1 gives an exemplary overview of the most severe hazards according to the Severity Class Scheme [27] that were part of the study. The participants ranked the estimated risk in a descending order. Since operational procedures and airports stay the same, the participants stated that possible hazards are most likely originated in limited visual and acoustic capabilities of the tower controller that is eager to direct and manage attention in respect to two dynamic and independent operational environments instead of one. The potential to miss or to confuse operational relevant information is considered as a consequence of inappropriate attentional management and the related lack of situational awareness. This is supported by findings, suggesting a link between low situational awareness and a an impaired detection of critical cues or failures in systems [28]. This can in turn lead to an overall decreased system performance and thus a poor detection performane of critical cues or failures in systems [29]. Complementary, the probability to confuse user controls of the instrument panel complies with the corresponding awareness for the location and functionality of technical equipment at the work position. In consequence, we consider an impaired situational awareness as a predominant causative factor of the hazards identified.

IV. Experimental Setup

A. Simulation Platform

For all trials, we used a simulation platform based on the software NARSIM from NLR (Figure 3) and provides:

- A video presentation with 14 simulated cameras with HD resolution using the ALICE visualization.
- Control of pan-tilt-zoom cameras via control stick and buttons with predefined zoom and direction of the camera.
- Instruments such as radar, flight strip and voice com for selecting frequencies and activating the telephone.

The single mode simulated Sundsvall Airport (ESNN) with a 360° horizontal view. The multi-mode included Sundsvall (right) and Örnsköldsvik (left) Airport with all screens following the airport-specific right-left split. The multi-mode featured a key limitation of a limited 180° FoV for each of the two airports. The instrument panel provided buttons for turning the FoV by about 102° to the right and left as well as a reset button that centered the FOV to the primary displays of the runway. The traffic was simulated by two pseudo pilots, controlling the aircraft and enabling realistic radio communication. Moreover, calls from the adjacent sector were simulated through a simulation operator (simop).

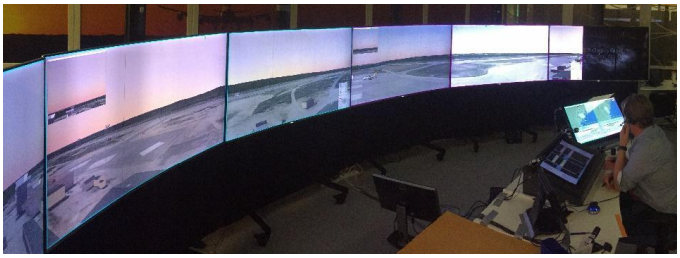

Figure 3. Tower Simulator NARSIM

A. Scenarios

Each test person was scheduled for four trials following a specific cross-over design of passing multi- and single-mode scenarios. The set of scenarios consisted of two multi-mode

scenarios (M1 and M2) and two single-mode scenarios (S1 and S2) with the characteristics outlined in Table 2. All scenarios shared the same qualitative composition of the scenario design. For mitigating any expectations of the test persons regarding timing and reappearance of tests, we disguised the pairing of the scenarios by randomized callsign and aircraft type.

.

TABLE I. HAZARDS AND RELATED TEST PROCEDURES

#	Hazard	SC	Safety Performance Indicator	Test Procedure(s)	Safety Metric
1.	Confusion of the Emergency Indicator Button in case of an accident	1	Awareness for the position of the correct inputs on the instrument panel and its associated airport	*Equipment handling*: Emergency Button Test	TTS HE
2.	Confusion of the Braking Action Values	1	Situational Awareness for the current braking action values	*SPAM:* Braking Action Value	TTS HE
3.	Confusion of visual conditions. Use of wrong runway holding positions not corresponding to the current conditions.	2	Situational Awareness for current QNH, Wind values and position of relevant objects	*SPAM:* QNH, Wind values, and position	TTS HE
4.	The limited FOV hides parts of the airport vicinity. No immediate visual contact to hidden objects possible.	4	Situational Awareness for the current position of objects in the CTR and on the maneuvering area	*Conflict Induction*: Moose and Car on the runway. Helicopter in the vicinity	TTD HE
5.	Confusion of frequencies (button or microphone). Landing clearance is given falsely.	3	Awareness for the current position of the correct inputs on the instrument panel	*Equipment handling*: Frequency Test	TTS HE
6.	Confusion of obstacles such as buildings or mountains in the environment of the Airport	1	Awareness for airport-related topological structure	No test available	
7.	Missing the transmission that ground vehicles vacated runway due to transmissions at both airports at a time	4	Situational Awareness for the current runway occupancy	*SPAM:* Position of snow sweeper	TTS HE
8.	Missing unknown movements on one airport while spending attention to the other airport	nA	Situational Awareness for all safety-relevant events on both airports	*Conflict Induction*: Moose and Car on the runway.	TTD HE
9.	Confusion of QNH value during landing situation	1	Situational Awareness for QNH	*SPAM:* QNH	TTS HE

B. Safety Performance Indicator

Based on the hazard list delivered from the safety workshop, safety performance indicators were developed for the hazard occurrences (Table 1). As highlighted in the section Hazard Identification, we defined two types of indicators:

- The error type "information confusion" originates in memory lapses. Distinctiveness is generally needed to avoid misattribution of the human memory that causes confusion about the origins of retrieved information [30]. We defined situational awareness for operational relevant information as an indicator of the ability to recall information from memory or to correctly locate information in the environment. The relation between human error types and SA model theory are explained more in detail in [9].
- Position awareness of the technical equipment of the work position including the position of user controls as well as the functionality of techniques and automation.

C. Test Procedures

The following three sections provide descriptions of our test procedures that were developed on the bases of the SPI. We distinguish three different test principles that are described in Figure 2. The allocation of the test procedure to the corresponding hazard and the related safety metrics are presented in Table 1.

1) Conflict Induction Tests (Table 3)

The appearances of objects in relevant areas are threat scenarios that aimed to test the test person's ability to detect and respond to any object that might be safety-relevant. The test was accomplished if the test person verbally comments "conflict". The object was removed to minimize the impact on simulated operations. We limited the test time to three minutes and counted an error in the case of missed detection.

2) Equipment Handling Secondary Task Test (Table 4)

The test person was instructed to follow the query as fast as possible. This tests the ability to find equipment and to identify the correct information while avoiding confusion. The test was accomplished if the test person has followed the instruction correctly.

3) Situation Present Assessment Method (SPAM) Query Test (Table 5)

The test person was called via the approach telephone and was asked to respond to a given question out of a set of predefined answers as fast as possible.

TABLE II. SCENARIO CHARACTERISTICS

Characteristic	Description
Length	90 min length for embedding all tests with fair distribution.
Weather	Varying weather and visibility conditions. The scenario starts with CAVOK and turns into a snowstorm after between 25-40 min with visibility about 600 m. The snowstorm fades out into a cloud cover without precipitation after another 15 min.
Traffic Mix	Eight scheduled movements with commercial traffic (six IFR and two VFR). Two flight school airplanes training touch and go exercises before and after the snowstorm.
Airports multi mode	Sundsvall and Örnsköldsvik with all movements fairly shared.
Airports single mode	Sundsvall in Single Mode
Scenario Pairs	The paired scenarios M1 and S1, as well as M2 and S2, are related to each other for supporting paired sampling tests of the safety metrics. This concerns especially the timing and the situational conditions of the test procedure execution. All other variables such as identification of aircraft and used aircraft models are randomized in its classes for covering up the systematic scenario frame concept and the recurring schedule of the embedded test sequence.
Metreport	Regular meteorological reports are provided every 30 min and three special meteorological reports are provided during the rapid change of pressure and temperature.
Braking Action	Braking action values are submitted via telephone.

TABLE III. CONFLICT INDUCTION TESTS

#	Test	Condition	Per trial
1.	Sudden appearance of a car on the runway.	During taxing of a/c on the runway	1
2.	Sudden appearance of a moose on the runway	During taxing of a/c on the runway	1
3.	Unauthorized entry of VFR into CTR with no flight plan filed.	Sim time scheduled	1

TABLE IV. EQUIPMENT HANDLING TEST

#	Test	Trigger Condition	Per Trial
4.	A helicopter appears suddenly on the backside. test person shall find it. This shall test the handling of the view field turn buttons.	During IFR on final in the CTR	1
5.	The test person shall set a certain frequency of the voice com system.	Sim time scheduled	2
6.	The Test person shall press the emergency button related to a given airport under control.	Sim time scheduled	2

D. Post Questionnaire

Besides the measurement of the reaction time and the success of the action, questionnaires were used for the following:

- Post-evaluation of the test procedures including aspects such as predictability, representativeness, and acceptance of the test procedures applied and the realism of the simulation.
- Self-evaluation measures, quantified by means of a self-report either before and after each experimental run.

TABLE V. SITUATIONAL AWARENESS QUERY TESTS

#	Test	Trigger Condition	Per Trial
7.	Wind Information	Sim time scheduled	2
8.	Location of a/c in the CTR in relation to the Tower (e.g. east-west)	Sim time scheduled	2
9.	QNH	Sim time scheduled	2
10.	Braking Action Values	Sim time scheduled	2
11.	Vehicle staying on the Runway	During snow sweeping, one vehicle stays on the runway while the other vacates.	1

V. RESULTS

A. Sample Characteristics

During a two-week period of simulations in September 2018, n = 8 test persons applied 32 trials in total. Test persons were either licensed or former-licensed tower controllers (five RTC Sundsvall, one Kristianstad, one Stockholm-Arlanda, one Linköping). Mean age of participants was 48.8 years (SD = 8.5), with one female and seven male controllers being part of the study. The participants had a mean work experience of 24.2 years (SD = 8.1). Six of the test person participated in the MERASSA first iteration pre-study [17] that was conducted on the same simulation platform, scenarios, and test procedures. This means that six out of eight test persons participated with an advanced level of training providing likely better performance than the other two. All test persons were trained in a one-hour training scenario in order to gain confidence with the instruments and the multi-mode specific functionalities of the system.

B. Reaction Times

The trials provided 492 reaction time samples from 544 scheduled tests. All reaction times are paired according to the paired scenarios as following:

$$\Delta RT = RT_{multi} - RT_{single}.$$

The resulting safety benchmark is visualized as box plots (Figure 4) showing the distribution of the RT paired samples. The related statistics are shown in Table 6. Any statistic Goodness of Fit test was negative regarding a possible normal distribution. We chose the Mann-Whitney U-Test instead that was applied with the two null-hypotheses:

- $\mathbf{H_0: RT_{Multi} \geq RT_{Single}}$ Test on significant accelerated work speed in multi-mode.
- $\mathbf{H_0: RT_{Multi} \leq RT_{Single}}$ Test on significant slower work speed in multi-mode.

The only statistical significance could be found for the emergency button test no. 6 in which the test persons needed on average 5.5 sec longer in multi-mode to find the correct button. The average value of two tests stands out indicating an accelerated reaction in multi-mode which is outlier caused.

The conflict induction and equipment handling tests show a comparable broad spreading of the samples whereas the SPAM tests show the densest spreading.

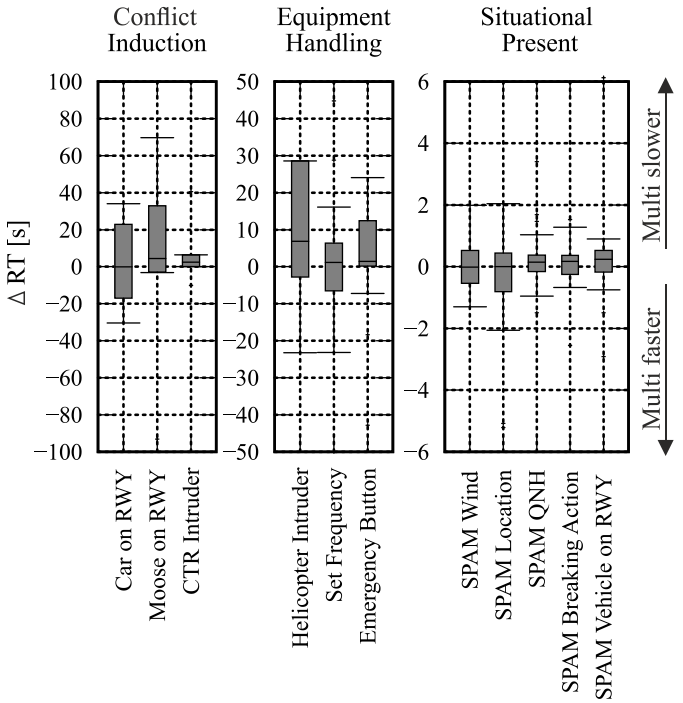

Figure 4. Safety Benchmark Plots

TABLE VI. SAFETY BENCHMARK STATISTICS

#	N	Paired Samples N_{Paired}	Mean [s]	Std [s]	P-value Paired U Test [%] H_0:$RT_{Multi} \geq RT_{Single}$	P-value Paired U Test [%] H_0:$RT_{Multi} \leq RT_{Single}$
1.	30	14	-6.37	39.9	52.0	50.0
2.	24	10	7.14	43.0	88.4	13.0
3.	19	6	7.00	17.4	86.0	20.9
4.	21	7	23.67	54.2	89.1	14.8
5.	61	29	4.54	30.0	54.2	46.6
6.	61	29	5.57	17.7	99.8	0.2
7.	62	30	0.01	0.8	48.8	52.1
8.	62	30	-0.50	2.2	48.0	52.9
9.	62	30	0.18	0.9	84.6	15.0
10.	59	27	0.11	0.8	87.6	12.9
11.	31	15	0.30	1.9	80.5	21.1

C. Human Error

The error list counts one wrong answer at the SPAM test concerning a QNH query in multi-mode.

D. Ex-Post Questionnaire

1) Post-Evaluation Test Procedures

The post-evaluation of the test procedure used a five-grade scale with 1-"completely disagree" and 5-"completely agree" by means of the questions described in Table 7. The results from 32 questionnaires are shown in Figure 5. The results show a clear picture for the questions 1-4 in which the impact of simulator artificiality, predictability of tests and any unfairness of the tests were considered as not significant. Question 5 had a spread over the scale in which test persons stated a significant impact on the attentional pattern in multi-mode. This is in line with the statements from [5] and [4] that highlighted similar problems of directing attention.

TABLE VII. QUESTIONS POST-EVALUATION

#	Question
1	I think that the artificiality of the simulation had an impacted on my behavior.
2	In general, I could predict the events more than in reality.
3	I prepared for the events because I could predict the occurrence
4	I'm of the opinion that the tests are not treating single and multi remote tower equally
5	I'm of the opinion that my attention was significantly impacted by the need to control two airports

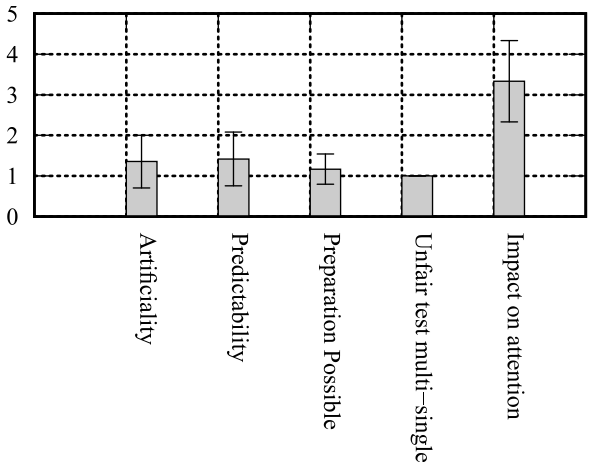

Figure 5. Post-Evaluation Test Procedure

2) Self-evaluation measures

The self-evaluation of performance during the simulation was measured after each trial by means of a questionnaire. The questionnaire consisted of items regarding the following variables: efficiency, safety, rule compliance, concentration demand, stressfulness and error amount. Participants were asked to estimate on a 5-grade scale with 1 (meaning low) to 5 (meaning high). Results are presented in Figure 6.

A Kolmogorov-Smirnov test indicated no normal distribution for efficiency ($D(32)=.413$, $p=.00$), safety ($D(32)=.370$, $p=.00$), rule compliance ($D(32)=.256$, $p=.00$), concentration demand ($D(32)=.247$, $p=.00$), stressfulness ($D(32)=.242$, $p=.00$) and error ($D(32)=.308$, $p=.00$). Hence, in order to analyse for differences between single- and multi-mode, a Mann-Whitney-U test was carried-out, which indicated no significant differences for any variable: Efficiency ($U=113$, $p=.45$), safety ($U=128$, $p=.99$), rule compliance ($U=126$, $p=.93$), concentration demand ($U=110.5$, $p=.48$), stressfulness ($U=110.5$, $p=.48$) and error ($U=121$, $p=.80$).

VI. DISCUSSION

A. Multi Remote Tower

A salient result is the extremely low number of human errors committed by the test persons. Compared to the number of errors committed in the previous iteration presented in [17] of 7 errors, a remarkable improvement of the overall work accuracy can be stated. As all other parameters of the simulation remain constant, the most reasonable explanation is the training and experience gained in both iterations, allowing for acting with higher confidence in multi- as well as single-mode. The higher confidence addresses both the speed and the accuracy similar to a decrease in task difficulty [31]. The other side of the coin is that safety analytics depends on sufficient sample sizes. The number of errors of this iteration is too low for any reliable statement based on this metric. Nevertheless, the reaction times provide sufficient samples for verifying the safety-relevance as it was planned in section two.

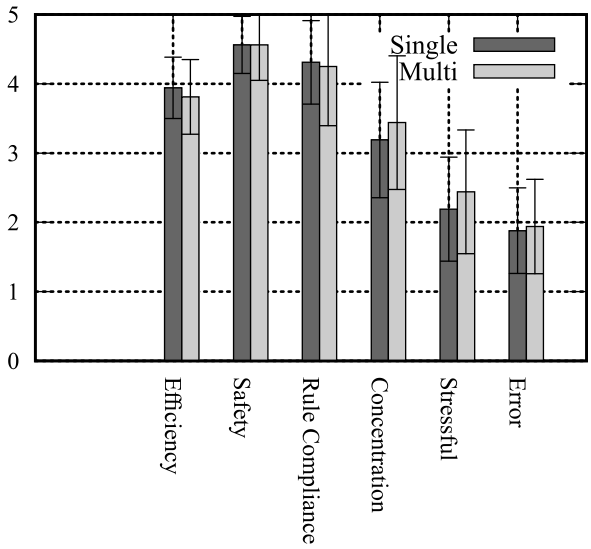

Figure 6. Self Evaluation

Another positive result is the fact that we could not find any positive test concerning H_0: $RT_{Multi} \geq RT_{Single}$ which indicates no verifiable tendency of the controllers to work faster in multi-mode. This corresponds to the low number of human errors that would have shown an increase in the case of accelerated work behavior. The diagrams Figure 4 comply with this tendency since the box plots are symmetric on the zero lines or even slightly positive shifted. We consider this observation as a sign of risk compensation in which the test persons regulate the work speed to a degree where they perceive the equal subjective risk in both modes. In this regards, an interesting case is the statistical significance of the emergency button test no. 6 that shows a clear slow down of the work speed in multi-mode. Since the emergency buttons are fairly accessible input controls in both modes, we consider this as a clear case of applied risk compensation. An alternative explanation are additional efforts and/or actions needed by the test person to access the same information in multi-mode. An example is the helicopter search test in which 23.7 s additional time in multi-mode can be clearly traced back to the use of buttons for turning the FoV. In these specific cases, clarification of the true origin was achieved using the video evidence that discovers these obvious actions.

The self-evaluation showed a slight deterioration for multi-mode for the responses errors, stress, concentration, rule compliance, and efficiency. This is complemented by the test procedure post-evaluation no. 5 that could not falsify the possible impact of multi-mode on the attention. Although we found no statistical significance, these responses share identical directions that might be caused by unfamiliarity of the multi-mode and the perceived uncertainty.

B. MERASSA

Generally, the self-evaluation and the test procedure post-evaluation are complying with the empiric results including both the reaction time and error rate. This is an indication of the validity of the results. Another sign is the equality of the empiric results that follows a common trend. The flexibility of the method allowed for developing a test procedure for any hazard using the work principles (Figure 2).

Summarizing the empirical and subjective response of the test persons, the results do not verify the safety-relevance of any hazard according to the ELoS criterion defined. The hazards might nevertheless be still of safety-relevance in operations because of the following reasons:

- *Simulator Induced Attention*: Test persons that act in simulators exhibit exceptional attention and energy that lies beyond non-routine behavior as they would have in real operations. Test results would be of increased authenticity and validity in a more acclimated environment.
- *Low Probability of Hazard Occurrence*: A long-term investigation, involving a large number of situations and diversity of conditions at the same time, would likely loom up human errors that lie beyond the observability of simulator-based studies due to statistic right censoring [32].

The MERASSA underlies the limitations of simulation-based evidence, especially at these two points. However, the use of the baseline and the iterative development of test procedures for sampling sufficient samples are the key factors to gain insight into the response of the test person to hazards.

VII. CONCLUSION

A methodological concept MERASSA has been presented that allows for the verification of the hazard's safety-relevance in a multi remote tower environment as part of the Eurocontrol SAM SSA. In this paper, we validated the method at the multi remote tower-case as a part of a safety assessment. Therefore, we identified hazards that are considered by experts as most threatening for operational safety. Test procedures and safety performance indicators have been defined that are suitable for testing the hazard's safety-relevance in multi-mode. An experimental design was developed including scenarios that embedded the tests in a simulation. Eight test persons executed multi- and single-mode scenarios in 32 trials, providing evidence for the verification of the hazard safety-relevance. The statistic test results show that the identified hazards cannot be verified. Rather, the results verified the test person's general

tendency to slow down or to maintain the same work speed across all tests applied. The most likely explanation of this behavior is a reaction of the test person to a perceived uncertainty and the need to compensate it. This might be time taken to double check information or other consciousness-raising activities that the test person considers as suitable in the respective test situation. The MERASSA method could be validated by comparing empirical and subjective data. At this stage of the multi remote tower, the empiric, as well as the subjective results, show in the entire response a lack of confidence in the novel concept of controlling two airports at a time. The impact on the habits and techniques how to manage attention is clearly identified as the predominant causal factor.

REFERENCES

[1] A. Oehme, R. Leitner, and N. Wittbrodt, "Challenges of Multiple Airport Control," in Aviation Psychology and Applied Human Factors vol. 3, pp 1-8, 2013.

[2] R. Leitner and A. Oehme, "Planning Remote Multi-Airport Control-Design and Evaluation of a Controller-Friendly Assistance System," in Virtual and Remote Control Tower, Cham, CH, Springer International Publishing Switzerland, 2016, pp. 139-160.

[3] ICAO, "Safety Management Manual (SMM) Doc. 9859 3rd Edition," ICAO, Montreal, CA, 2013.

[4] A. Papenfuss and M. Friedrich, "Head up only - a design concept to enable multiple remote tower operations," in 35th Digital Avionics Systems Conference, Sacramento, USA, 2016.

[5] P. Kearney and W. Li, "Multiple remote tower for Single European Sky: The evolution from initial operational concept to regulatory approved implementation," in Transportation Research Part A: Policy and Practice, pp. 15-30, 10 2018.

[6] B. Josefsson, J. Jakobi, A. Papenfuss, T. Polishchuk, C. Schmidt, and S. L, "Identification of Complexity Factors for Remote Towers," in 8th SESAR Innovation Days, Salzburg, AT, 2018.

[7] J. Vogt, T. Hagemann, and M. Kastner, "The impact of workload on heart rate and blood pressure in en-route and tower air traffic control," Journal of psychophysiology, pp. 297-314, 2006.

[8] B. Josefsson, C. Chalon-Morgan, and E. Pinska-Chauvin, "Remotely provided Air Traffic Services for two low density aerodromes - Appendix F: HP Assessment Report," SESAR SJU, Brussels, BE, 2015.

[9] L. Meyer and H. Fricke, "Investigating the Safety-Relevance of Limited Distinctive Features on a Multi Remote Tower-Working Position," in 6th SESAR Innovation Days, Delft, NL, 2016.

[10] S. Shapell and D. Wiegmann, "U.S. naval aviation misshaps 1977-92: Differences between single- and dual-piloted aircraft," in Aviation, Space, and Environmental Medicine, pp. 65-69, 1996.

[11] M. Llobet Lopez, "OFA06.03.01 Remote Tower - Safety Assessment Report for Multiple Remote Tower," SESAR JU, Brussels, BE, 2015.

[12] P16.06.01, "Accident Incident Models in MS Visio – AIM V10-3,," SESAR JU, Brussels, BE, 2015.

[13] M. Llobet Lopez, "Results from the Safety Questionnaire," SESAR SJU, Brussels, BE, 2015.

[14] Eurocontrol, "Safety Case Development Manual," Eurocontrol, Brussels, BE, 2006.

[15] Eurocontrol, "SAM SSR Safety Assurance and Evidence Collection," Brussels, BE, 2003.

[16] L. Meyer, J. Lundberg, B. Josefsson, L. Danielson, and H. Fricke, "Methods for the Empiric Risk Assessment of Socio-technical Systems in ATM," in 5th SESAR Innovation Days, Bolognia, IT, 2015.

[17] L. Meyer, B. Josefsson, M. Peukert, and J. Lundberg, "An Empiric Stress Test Validation for Multi Remote Tower Safety Assessment," in SESAR Innovation Days, Salzburg, 2018.

[18] SAM-TF, "Air Navigation System Safety Assessment Methodology," Eurocontrol, Brussels, BE, 2004.

[19] H. Bubb, „Human Reliability: A key to improved quality in manufacturing," Human Factors and Ergonomics in Manufacturing & Service Industries vol. 15 no. 4, pp. 353-368, 2005.

[20] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," in Journal of experimental psychology vol. 47 no. 6, p. 262-269, 1954.

[21] B. U. Forstmann, R. Ratcliff, and E.-J. Wagenmakers, "Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions," Annual Review of Psychology , pp. 641-666, 2016.

[22] A. Vuckovic, P. J. Kwantes, and A. Neal, "Adaptive decision making in a dynamic environment: A test of a sequential sampling model of relative judgment," Journal of Experimental Psychology Vol. 19 no.3, p. 266–284, 2013.

[23] A. Vuckovic, P. J. Kwantes, M. Humphreys, and A. Neal, "A Sequential Sampling Account of Response Bias and Speed–Accuracy," Journal of Experimental Psychology vol 20 no. 1, p. 55–68, 2014.

[24] G. J. Wilde, "Risk homeostasis theory: an overview," Injury prevention 4.2, pp. 89-91, 1998.

[25] M. G. Fennema and D. N. Kleinmuntz, "Anticipations of effort and accuracy in multiattribute choice," Organizational behavior and human decision processes vol 63 no.1, pp. 21-32, 1995.

[26] H. de Jong, "Guidelines for the identification of hazards: how to make unimaginable hazards imaginable?," NLR, Amsterdam, NL, 2004.

[27] The European Commission, "Commission regulation (EC) No 1035/2011 - laying down common requirements for the provision of air navigation services," Official Journal of the European Union, pp. 23-41, 2011.

[28] N. Stantona and J. Piggott, "Situational awareness and safety 30(3)," Safety Science, pp. 189-204, 12 2001.

[29] D. Kaber and M. Endsley, "Out- of- the- loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety," Process Safety Progress vol. 16 no. 3, pp. 126-131, 2004.

[30] D. L. Schacter, and A. L. Wiseman, "Reducing memory errors: The distinctiveness heuristic," in Distinctiveness and memory, New York, NY, Oxford University Press, 2006, pp. 89-107.

[31] D. Standage, D.-H. Wang, R. P. Heitz, and P. Simen, "Toward a unified view of the speed-accuracy trade-off," Frontiers in neuroscience, vol. 9, p. 139, 2015.

[32] L. Meyer, K. Gaunitz, and H. Fricke, "Experimental Study for the Empiric Risk Analysis of Sociotechnical Systems in ATM," in Human Factors in Transportation, Boca Raton, USA, Taylor Francis Group, CRC Press, 2016, p. 313–329.

[33] ICAO, "Annex 19 - Safety Management 1st Edition," ICAO, Montreal, CA, 2013b.

Biographies

Lothar Meyer (1981) is a safety engineer at Luftfartsverket (LFV), the Swedish Air Navigation Service Provider. He holds a degree in electrical engineering and a doctorate in air traffic services from the Technische Universität Dresden, where he worked as a research associate in the field of aviation safety. His areas of expertise covers socio-technical systems, risk assessment and airport surveillance technologies

Maximilian Peukert (1991) received a MSc degree in psychology and human performance in 2017. After working as a research associate and lecturer at the chair of engineering psychology at Technische Universität Dresden, he is since 2018 active as a Human Performance Specialist for LFV Research & Innovation.

Jonas Lundberg (1974) has a PhD in Computer Science from Linköping university, 2005, in the area of Interaction Design, with Human-Automation Collaboration as the current main area of research. He has also conducted research in Safety Science since 2006. Since 2016 he is associate Professor in Information Design at Linköping University, Sweden.

Billy Josefsson (1962) is a senior Air Traffic Controller with a background in psychology. Since 1994 active within research and development in human performance, safety and human factors worldwide. Since 2014 he is Manager for Automation and Human Performance at LFV Research & Innovation