

FSFT- Flygsäkerhet Fjärrstyrda Torn Slutrapport

REVISIONSHISTORIK

VERSION	DATUM	UPPRÄTTAD AV	KOMMENTAR
0.1	2019-05-08	Lothar Meyer	Initial
0.2	Maj 2019		Minor updates
0.3	2019		Minor updates related to comments received

INNEHÅLL

REVISIONSHISTORIK.....	2
1 SAMMANFATTNING	4
2 INTRODUCTION.....	4
3 THE MERASSA CONCEPT.....	8
3.1 Safety Metrics.....	9
3.2 Safety Benchmark.....	9
3.3 Verification Criteria for the ELoS.....	10
4 HAZARD IDENTIFICATION	13
5 EXPERIMENTAL SETUP.....	14
5.1 Simulation Platform.....	14
5.2 Scenarios	15
5.3 Safety Performance Indicator	16
5.4 Test Procedures.....	16
5.5 Post Questionnaire.....	17
6 RESULTS	18
6.1 Sample Characteristics	18
6.2 Reaction Times	18
6.3 Human Error	19
6.4 Ex-Post Questionnaire	20
7 DISCUSSION	22
7.1 Multi Remote Tower	22
7.2 MERASSA	23
8 CONCLUSION.....	24
9 MILESTONES	24
9.1 Experience of the project participants.....	25
10 REFERENCES.....	25

1

SAMMANFATTNING

Projektets syfte var att utveckla en metod som skapar bevis för en acceptabel flygsäkerhetsnivå i balans med kapacitetsbehov om man samtidigt fjärrstyrer två torn med en flygledare (Multi Remote Tower). Metoden kallas även MERASSA och omfattar objektiva och subjektiva bevis genom att testa konceptet i en tornsimulator och jämföra resultat med ett single baseline scenario som testas på samma sätt. Testprocedurer består av konflikt, "secondary task" och situationsmedvetenhetstester som levererar reaktionstider och framgång av uppgifter som testpersonen utövar i simulatoren på ett kontrollerat sett. Statistiska jämförelser skapar ett "safety benchmark" som visar förhållandet mellan arbetshastighet och felfrekvenser som är specifika och karakteristiska för testpersonens arbets säkerhet i en ny miljö som multi remote tower.

Tre flygledare och en pilot diskuterade och identifierade under två safety workshops riskkällor som har sina ursprung i mänskliga fel medan de kontrollerar två flygplatser samtidigt. Resultat av dessa safety workshops var en lista av riskkällor samt allvarlighetsgraden som visar en större riskpotential i förväxling av operativt relevant information tex. QNH, vindförhållande och bromsverkan.

Metodiska koncept testades med hjälp av experimentella valideringar med NARSIM tornsimulatorens åtta flygledare som testpersoner. Första iterationen var en förstudie för att kunna utveckla, förstå och förbättra metodiska koncept. Andra iterationen visar konceptets förmåga att skapa safety benchmarks och verifiera riskkällor från safety workshopen. Sammanfattningen var att flygledare inte visade någon förhöjd potential för mänskliga fel men däremot fanns tendensen att testpersonen arbetar långsammare i multi mode. Förklaringen kan vara att testpersonerna kände sig osäkra och mindre trygga i multi mode arbetsområdet. Att arbeta långsamt är ett känt beteendefenomen som kallas Riskhomeostas efter Wilde samt en bra indikator för säkerhetsrelevansen av en potentiell riskkälla. Metoden som utvecklades inom projektet visar förmågan att mäta mänskliga prestanda och evaluerar en balans mellan säkerhet (mänskliga fel) och kapaciteten (arbetshastighet).

2 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, it is rational to assess the safety of the concept. However, current safety assessment methodology relies on subjective statements of experts and does not evaluate if the hazards found 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 within the safety case for the approval process. 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. We will use the LFV concept as the basis for our discussion.

In single runway remote tower services, LFV has had the focus of designing the work position to resemble conventional towers as close as possible, preserving the well-established work methods of the tower controllers to a large extent. 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 is emerged and a 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 support intensifies the symbiosis between the 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 significant 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 undertaken 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 aircrafts and increased workload as a result of simultaneous tasks were found. 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. Thus, these emphasize the critical role of the tower controller to keep track of all operational relevant information in situations where traffic situations and related conditions changes 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 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 feed back 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

lift the awareness of safety workshops participants for multi remote tower related risks to a superior level.

Following the task definition of the Eurocontrol SAM SSA¹ [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. By this, the demand for safety and capacity can be balanced in a mainly objective manner. 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 . 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 form of a hazard list with the estimated degree of severity are shown. The experimental setup up 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 distribution and appropriate diagrams. Finally, conclusions about the multi remote tower hazards are finally drawn on the empiric as well as the subjective data collected.

3 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 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 steps in 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 indicators² (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 a 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 feed back 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 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 assume the control of all other variables in regards of the actual operational conditions under which the test is applied. This concerns especially the timing of execution and the actual traffic situation.

3.1 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 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 variables in the ATC context indicating human uncertainty and work performance.

We distinguish reaction time 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 potential 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.

3.2 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.

3.3 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 use risk compensation theory. According to that, the addition or removal of safety factors will result in a behavioural change that leads to compensatory measures involving adjustments to work speed and accuracy [24]. According to that theory, 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 pressures [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.

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 equal number of human errors) indicates the response of the test persons to an increased difficulty and uncertainty in multi-mode. From the 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.

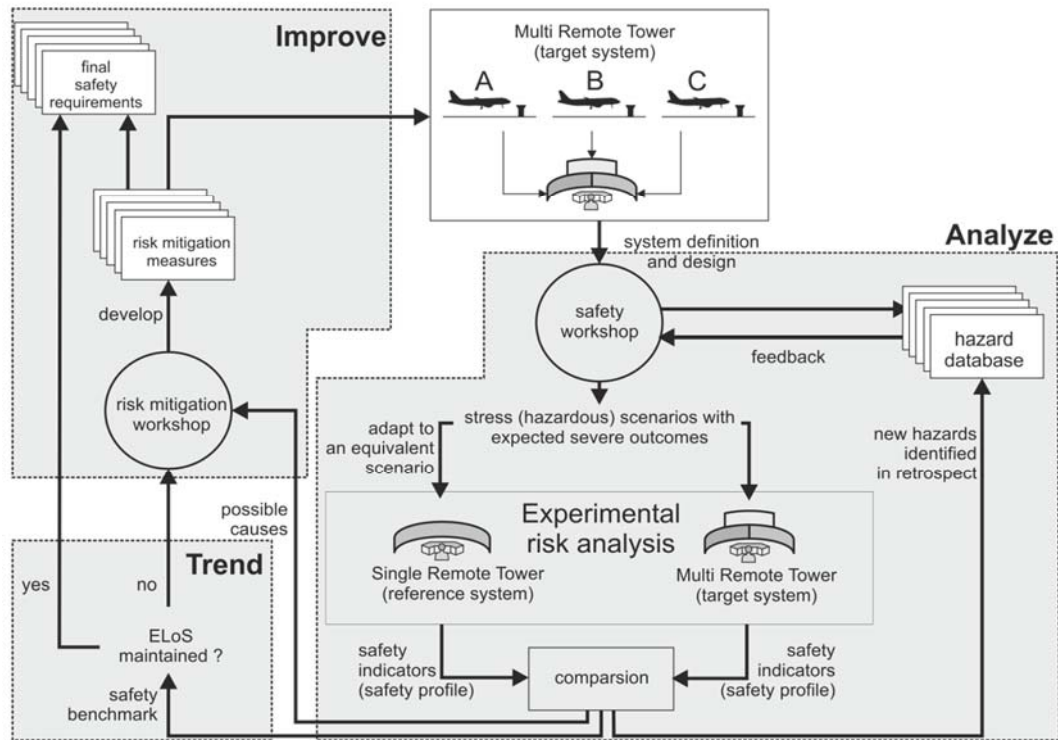
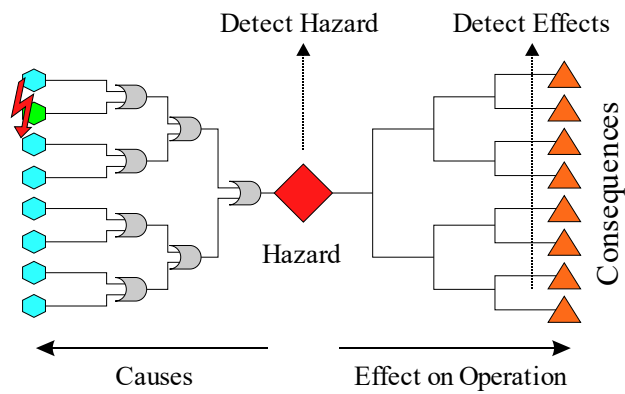
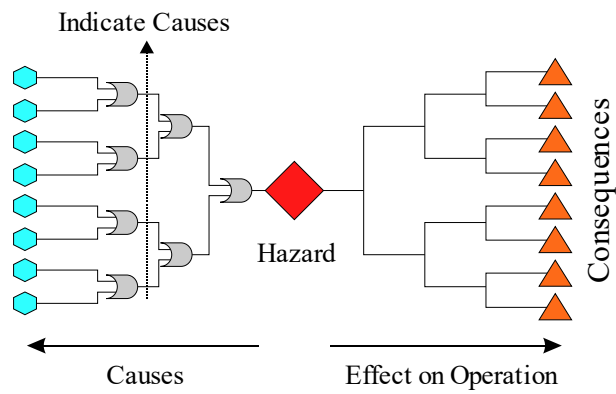


Figure 1: MERASSA Concept

Conflict Induction Test



Query Test



Secondary Task Test

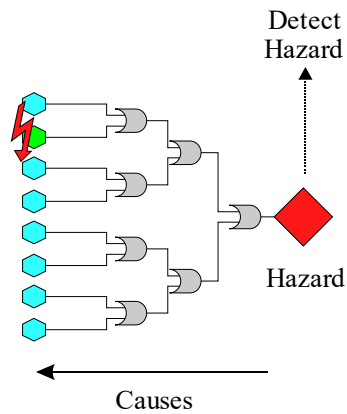


Figure 2: Test Procedures Working Principle

4 HAZARD IDENTIFICATION

By means of a three-day safety workshop, three tower controllers from Stockholm-Arlanda, Stockholm-Bromma and Östersund as well as a A320-rated Pilot identified a number of multi-mode-related hazards. The workshop was conducted in accordance to [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.

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 multi-mode-specific 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 rather to confuse operational relevant information is considered as a consequence of an inappropriate attentional management and the related lack situational awareness. These statements are supported by findings such as the loss of situational awareness can lead to an overall decreased system performance [28]. A deterioration of situational awareness can lead to an impaired detection of critical cues or failures in systems [29]. Complementary, the probability to confuse user controls of the instrument panel complies with the corresponding position awareness for the technical equipment of the work position. In consequence, we consider an impaired situational awareness as a predominant causative factor of the hazards identified.

Table 1: 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	<i>Equipment handling:</i> Emergency Button Test	TTS HE
2.	Confusion of the Braking Action Values	1	Situational Awareness for the current braking action values	<i>SPAM:</i> Braking Action Value	TTS HE
3.	Confusion of visual conditions. Use of wrong runway holding	2	Situational Awareness for current QNH, Wind values	<i>SPAM:</i> QNH, Wind values and position	TTS HE

	positions not corresponding to the current conditions.		and position of relevant objects		
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	<i>Conflict Induction:</i> Moose and Car on runway. Helicopter in the vicinity	TTD HE
5.	Confusion of frequencies (button or microphone). Landing clearance given falsely.	3	Awareness for the current position of the correct inputs on the instrument panel	<i>Equipment handling:</i> 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	<i>SPAM:</i> Position of snowsweeper	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	<i>Conflict Induction:</i> Moose and Car on runway.	TTD HE
9.	Confusion of QNH value during landing situation	1	Situational Awareness for QNH	<i>SPAM:</i> QNH	TTS HE

5 EXPERIMENTAL SETUP

5.1 Simulation Platform

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

- A video presentation with 14 simulated cameras with HD resolution using the ALISE 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 simulates Sundsvall Airport (ESNN) with a 360° horizontal view. The multi-mode simulates Sundsvall (right) and Örnsköldsvik (left) Airport with all screens following the airport specific right-left split. The multi-mode features a key limitation of a limited 180° FoV for each of the two airports. The instrument panel offers buttons for turning the FOV by about 102° to the right and left as well as a reset button that centers the FOV to the primary displays of the runway.

Traffic is simulated by two pseudo pilots, controlling the aircraft and enabling realistic radio communication. Moreover, calls from the adjoining sector were simulated through a simulation operator (simop). This role contains the tasks to support the pseudo pilots by implementation of aircraft in the simulation system according to the schedule.

Additionally, the simop monitored the simulator and ensured a proper simulator performance in all trials.

5.2 Scenarios

Each test person is scheduled for four trials following a specific cross-over design of passing multi- and single-mode scenarios. The set of scenarios consists 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 share same qualitative composition of the scenario design. For mitigating the expectations of the test persons in terms of the timing and the appearance of any test, we disguise the pairing of the scenarios by randomized callsign and aircraft type and thus confusing the test persons.



Figure 3: Tower Simulator NARSIM

Table 2: Scenario Characteristics

CHARACTERISTIC	DESCRIPTION
Length	90 min length for embedding all tests with a 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 closed cloud cover without precipitation after another 15 min.
Traffic Mix	8 schedules movements with commercial traffic (6 IFR and 2 VFR). 2 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 aircrafts and used aircraft models are randomized in its classes for covering up the systematic scenario frame concept and the reoccurring schedule of the embedded test sequence.
Metreport	Regular meteorological reports are provided every 30 min and three special meteorological reports are provided during rapid change of pressure and temperature.
Braking Action	Braking action values are submitted via telephone

5.3 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 define 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 rationale of connecting 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 functionality of techniques and automation.

5.4 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.

5.4.1 Conflict Induction Tests (Table 3)

The appearance of objects in relevant areas are threat scenarios that shall test the test person’s ability to detect and respond to any object that might be safety-relevant. The test is accomplished if the subject verbally comments “conflict”. The object disappears immediately for minimizing the impact on simulated operations. We limited the test time to three minutes and counted an error in the case of missed detection.

5.4.2 Equipment Handling Secondary Task Test (Table 4)

Test persons were instructed to follow the query as fast as possible. We aim to test the ability to find equipment and to identify the correct information while avoiding confusion. The test is accomplished if the test person has followed the instruction correctly.

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

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

Table 3: 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 flightplan filed.	Sim time scheduled	1

Table 4: 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

Table 5: Situational Awareness Queries 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.	Breaking 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

5.5

Post Questionnaire

Beside the measurement of the reaction time and the success of 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.

6 RESULTS

6.1 Sample Characteristics

During a two-week period of simulations, $n = 8$ 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 exact same simulation platform, scenarios and test procedures. This means that six out of eight test persons participated with an advanced level of training providing perhaps better performance than the other two. The test persons were trained in a one-hour training scenario gaining confidence with the instruments and the multi-mode specific functionality.

6.2 Reaction Times

The trials provided 492 reaction time samples from 544 scheduled tests. Missing samples were caused by formal errors and lapses in the thorough execution of test procedures. 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:

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

6.3 Human Error

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

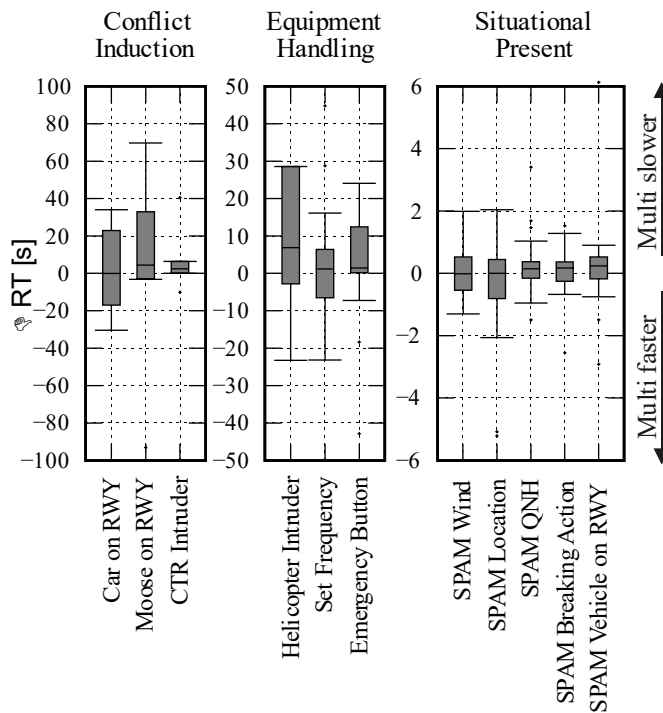


Figure 4: Safety Benchmark Plots

Table 6: Safety Benchmark Statistics

#	N	Paired Samples N _{Paired}	Mean [s]	Std [s]	P-value Paired U Test [%] H ₀ :RT _{Multi} ≥ RT _{Single}	P-value Paired U Test [%] H ₀ :RT _{Multi} ≤ 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

6.4 Ex-Post Questionnaire

6.4.1 Test Procedure Validation

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 is 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 was considered as not significant. The 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.

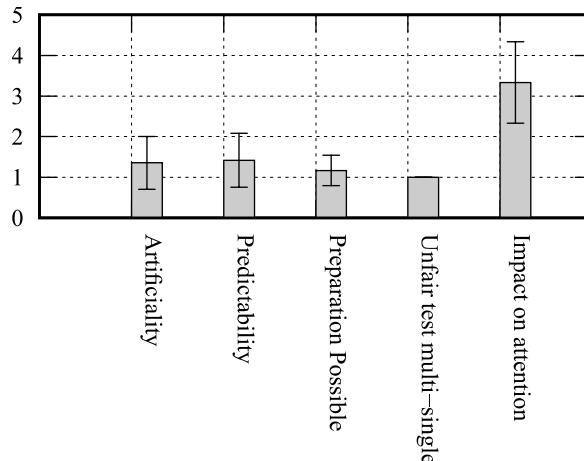


Figure 5: Post-Evaluation Test Procedure

6.4.2 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 ($U=121$, $p=.80$).

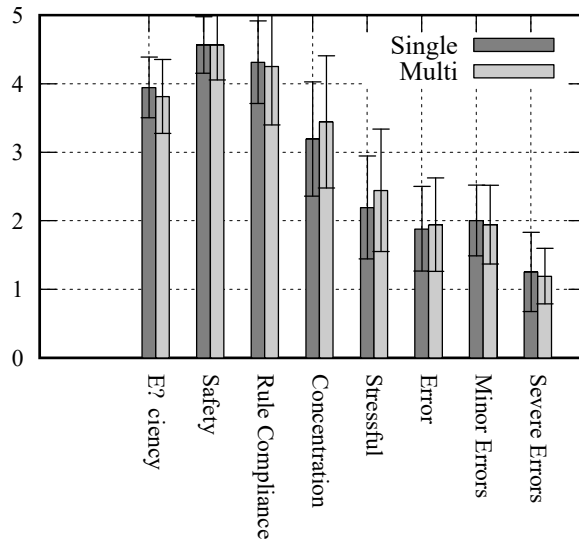


Figure 6: Self Evaluation

7 DISCUSSION

7.1 Multi Remote Tower

A salient result is the extreme 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 of 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.

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 line 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 is the need for additional efforts and/or actions 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 this 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.

7.2

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 for 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 the insight into the response of the test person to hazards.

8 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 while balancing safety and capacity demands objectively. Therefore, we identified hazards that are considered by experts as most threatening for the 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.

MERASSA shows the capability of balancing the demand between safety (error rate) and capacity (work speed) in an objective manner by using the safety benchmark for optimizing the operational concept while maintaining an acceptable level of safety at all steps of the system development.

9 MILESTONES

- 1st Safety Workshop with 3 tower controllers, identifying hazards and related risks by means of 3 tower controllers and 1 Airbus A320 captain from Novair
- Developing the experimental design and test procedure concept. Implementation to the NARSIM Simulator in the ATCC Sturup
- Training and education of the test person in the simulator including the test execution

- Execution of the main trials by means of 6 tower controllers from RTC Sundsvall, Arlanda and Kristianstad
- Risk mitigation workshop in Norrköping including 5 persons involved in the development of multi remote tower
- 2nd Safety Workshop by means of 2 tower controllers and 1 Airbus A320 captain from Novair in Stockholm
- Execution of the 2nd iteration trials with 8 test persons from RTC Sundsvall, Arlanda and Kristianstad

The project and the writing of this report were delayed compared to the original plan, this was mainly related to the availability of resources and the simulator platform.

10 EXPERIENCE OF THE PROJECT PARTICIPANTS

The project participants considered the preparation and the execution of the trials a major bottleneck of the whole project. The efforts needed in terms of time, skills and measures to account for uncertainty in the planning of the simulation platform was regarded as intensive but manageable. Experience was collected on how to handle the simulation equipment and the recordings of the data sources, including video, eye tracking and workload. The data collected from the trials was considered as highly valuable since many open conclusions could be drawn beyond of the principle research hypothesis. The data was shared with projects such as RESKILL (Project funded by Trafikverket) for developing tools to analyze the behavior of the tower controller.

11 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. Chalou-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 mishaps 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, Bologna, 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. Stanton 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.

12

ABBREVIATIONS

ATC: Air Traffic Control

ATM: Air traffic management

CAVOK: Ceiling and visibility OK

CTR: Control Zone

EloS: Equivalent Level of Safety

E-OCVM: European Operational Concept Validation Methodology

FHA: Functional Hazard Assessment

IFR: Instrument Flight Rules

MERASSA: Methods for the Empiric Risk Assessment of Socio-Technical Systems in ATM

NARSIM: Forsknings- och utvecklingssimulator som används på LFV.

NLR: Royal Netherlands Aerospace Centre

QNH: Altimeter sub-scale setting to obtain elevation when on the ground (ICAO definition)

RTC: Remote tower center

SA: Safety Assessment

SAT: Speed-Accuracy-Tradeoff

SESAR: Single European Sky ATM Research

SIMOP: simulation operator

SPAM: Situation Present Assessment Method

SPI: Safety performance indicators

TTD: Time-To-Detect

TTS: Time-To-Solve

VFR: Visual Flying Rules