D6.4: VALIDATION REPORT EUROPEAN CONSORTIUM (FINAL ITERATION)

SECURITY: PUBLIC

Lead beneficiary: HungaroControl

Contractual Due Date: M42

Actual Submission Date:
30/06/2023 (M42)

Grant Agreement number:	875154
Project acronym:	GREAT
Project title:	GREENER AIR TRAFFIC OPERATIONS
Funding scheme:	RIA/ H2020
Start date of the project:	January 1st, 2020
Duration:	42 months
Project coordinator (organisation):	Michael Finke (DLR)
Phone:	+49 531 295-2921
E-mail:	Michael.Finke@dlr.de
Project website address:	www.project-great.eu

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 875154 GreAT.

DOCUMENT INFORMATION

DOCUMENT NAME	D6.4: Validation Report Final iteration (European Consortium)
VERSION	VF
VERSION DATE	30/06/2023
AUTHOR	Fanni Kling (HC), Attila Pásztor (HC), Kristóf Kovács (HC), Thorsten Mühlhausen (DLR), Àlex Ramonjoan (Pildo), Rabeb Abdellaoui (DLR), Hilke Boumann (DLR), Nils Ahrenhold (DLR), Izabela Stasicka (DLR), Lukas Tyburzy (DLR), Marco-Michael Temme (DLR)
SECURITY	Public

DOCUMENT APPROVALS

	NAME	ORGANISATION	DATE
COORDINATOR	Michael Finke P/O Marco-Michael Temme	DLR	30/06/2023
WP LEADER	Michael Finke P/O Marco-Michael Temme	DLR	30/06/2023
TASK LEADER			
OTHER (QUALITY)	Jetta Keranen	L-UP	29/06/2023

DOCUMENT HISTORY AND LIST OF AUTHORS

VERSION	DATE	MODIFICATION	NAME (ORGANISATION)
V0.1	23/09/2022	First draft	Rabeb Abdellaoui (DLR)
V0.2	12/06/2023	Integration of contributions	Fanni Kling (HC), Kristóf Kovács (HC), Attila Pásztor (HC)
V0.3	15/06/2023	Contribution of diagrams, results, formatting, validation	Marco Temme (DLR), Àlex Ramonjoan (Pildo)
VF	30/06/2022	Validation, quality control	Marco Temme (DLR), Jetta Keranen (L-UP)

DISTRIBUTION LIST

FULL NAME OR GROUP
GreAT Consortium EU
European Commission / CINEA
All public

EXECUTIVE SUMMARY

Addressing environmental challenges, especially global warming, is more than ever a must for the international community. This matter is becoming an increasing priority at regional and global level. Europe has made commitments to reduce the aviation's environment footprint. Hence, it is contributing to climate change, increasing noise, affecting local air quality and consequently affecting the health and quality of life of European citizens. Due to Covid-19, the air traffic was drastically reduced for more than two years and it is expected that it will need five to ten years to recover to 2019 numbers. This offers the chance to rebuild it greener than before. The air traffic in Europe was growing until 2019 and is expected to continue increasing significantly in the future again in order to cope with the growing demand for mobility and connectivity. A long-term effect on the environment from aviation sector, mainly caused by aircraft noise and exhaust gases (especially CO_2, nitrogen oxides NOx and methane), make it a clear target for mitigation efforts. The future growth of aviation shall go hand in hand with environment sustainability policies. Therefore, studies and research are being conducted in Europe exploring possible optimization of the aircraft technologies as well as Air Traffic Management operations. Given the close interdependency between flight routing and environment impact, optimization in flight trajectory design and ATC operations are an appropriate means to reduce the emissions in short- and medium-term periods.

The international project “Greener Air Traffic Operations” (GreAT) has been launched in line with this perspective. This project is conducted in cooperation between Chinese and European partners.

With this present document, European partners intend to take the work started in the framework of MWP2, 3 and 4 to the next level. In MWP2, the theoretical basis was laid down, and project partners gained an insight view into the specific characteristics of each other's ATM system. This theoretical basis can be found in D2.1 [Finke 2021]. The developments indicated in this said document were further elaborated in the framework of MWP3 and MWP4, and are validated within the MWP6. The performed validation activities aimed to assess whether the GreAT green concept elements and the identified solutions are able to live up to the expectations of the consortium members. The results presented within this document are further evaluated from environmental impact point of view under MWP7.

The validation report gathers the results of the validation exercises performed by DLR, HC and Pildo Labs. In line with the Validation Plan D6.1 [Kling 2021], all parties assessed a common set of KPAs and validation objectives despite the diversity of ATM system developments under the umbrella of the GreAT project. GreAT was an environmental focused project, therefore environment is the most important Key Performance Area (KPA), and with one exception, is applied to all system developments. Additionally, the project investigated also the important KPAs safety, capacity, efficiency and cost-efficiency, and also Human Factors, where applicable, according to the specific characteristics of each system. On the Chinese side, a validation report was also produced, which summarizes and evaluates the relevant project results from the Chinese developments. This is based on the same document structure as well as validation objectives and was decided to enable a common understanding and framework on both continents. At the same time, the two reports respect the respective practices and systems' special characteristics when carried out the validation activities.

DISCLAIMER

The information, documentation and figures in this document are written by the GreAT consortium under EC grant agreement no. 875154 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Acronym	Signification
4D-CARMA	4-Dimensional Cooperative Arrival Manager
ACACIA	Advancing the Science for Aviation and Climate
ADCO	Arrival Departure Coordinator
A-FMS	Advanced Flight Management System
AI	Artificial Intelligence
ALTERNATE	Assessment of alternative aviation fuels development
AMAN	Arrival Manager
ANSP	Air Navigation Service Provider
APP	Approach Control
ARR	Arrival
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
ATMOS	Air Traffic Management and Operation Simulator
CAAC	Civil Aviation Authority China
CADEO	Controller Assistance for Departure Optimization
CCO	Continuous Climb Operations
CDA	Continuous Descent Approach
CDO	Continuous Descent Operations
CLIMOP	Climate assessment of innovative mitigation strategies towards operational improvements in aviation
CONOPS	Concept of Operations
CWP	Controller Working Position
DMAN	Departure Manager
EASA	European Union Aviation Safety Agency
EC	Executive Controller
EDDM	Munich Franz-Josef Strauß Airport

Acronym	Signification
EDDH	Hamburg airport
EFCA	Early Full Clearance Approach
EUROCAE	European Organisation for Civil Aviation Equipment
E-OCVM	European Operational Concept Validation Methodology
ETA	Expected Time of Arrival
FISA	Fatigue Instantaneous Self-Assessment
FMS	Flight Management System
FRA	Free-route Airspace
ft	Feet
HitL	Human-in-the-Loop
HMI	Human Machine Interface
IAF	Initial Approach Fix
IBM	International Business Machine
ICAO	International Civil Aviation Organization
ISA	Instantaneous Self-Assessment
KPA	Key Performance Area
KPI	Key Performance Indicator
LHBP	Budapest Liszt Ferenc International Airport
LMP	Late Merging Point
MWP	Main Work Package
NARSIM	NLRs Air traffic management Real-time Simulator
NM	Nautical Mile
PC	Planner Controller
SASHA-Q	Situation Awareness for SHAPE Questionnaire
SATI	SHAPE Automation Trust Index
SESAR	Single European Sky ATM Research Programme
SHAPE	Solution for Human-Automation Partnerships in European ATM
SID	Standard Instrument Departure

Acronym	Signification
SMAN	Surface Manager
SOBT	Scheduled Off-block Time
SPSS	Statistical Package For The Social Sciences
STAR	Standard Arrival Route
SUS	System Usability Scale
RTS	Real-time Strategy
TLX	NASA Task Load Index
TMA	Terminal Manoeuvring Area
TMAN	Turn-around Manager
TOBT	Target Off-block Time
TRA	Temporary Reserved Area (Military Reserved Airspace)
TRACC	Taxi Routes for Aircraft: Creation and Controlling
VALP	Validation Plan
VALR	Validation Report

1. INTRODUCTION

1.1. PURPOSE OF THE DOCUMENT

This document provides the Validation Report for GreAT project. It summarizes the validation exercises previously defined in the validation plan [Kling 2021], describes how they have been conducted and provides preliminary analyses, conclusions, recommendation as well as potential next steps.

The Validation Report conveys the overall series of validation activities with the aim of delivering results that may contribute to the successful implementation of GreAT concept elements in the future.

1.2. SCOPE

This document presents the validation activities performed by the European partners in the scope of GreAT project as well as the related results and recommendations. These activities are part of the MWP6 who's aim was to bring together the various performance areas and validate the developed airspace design and ATC support tools, and thus provide feedback on the overall concept applicability.

The results related to the validation activities performed by the Chinese partners is addressed in separate document.

1.3. INTENDED LEADERSHIP

This section describes the intended audience for this document. In general, readers of this document can be:

1) Readers internal to the project, using this document as input for their own activities.

2) Readers of GreAT sister projects (ACACIA, CLIMOP, ALTERNATE), using to follow latest developments and approaches, and to drive scientific exchange between the sister projects. This is for aligning the activities of all four projects and identifying synergy effects. Finally, this document can also serve as reference for scientific publications.

3) Readers from the GreAT Advisory board, in order to provide input and to follow the developments from a stakeholder point of view.

4) Readers involved in current and future projects dealing with reducing the impact of aviation on climate change and other environmental parameters, especially to build upon the approaches described in this document; and to align other developments (e.g. modifications to aircraft propulsion and airframe) with it.

5) Readers from air navigation service providers or other stakeholders not involved in the project but effected from its developments (especially airports, airlines or ATC equipment providers).

6) Standardization bodies and regulating authorities and organizations like ICAO, EASA, EUROCONTROL or CAAC.

All other interested members of aviation community.

1.4. STRUCTURE OF THE DOCUMENT

This document contains the following sections:

Chapter 1 Introduction – describes the purpose and scope of the document, the intended audience and the document structure.

Chapter 2 Context of the Validation – provides of a short summary of the validation plan. It presents the evaluated concept, the validation approach, enclosing the validation objectives, success criteria and the aims and techniques needed to conduct the validation exercises. It also includes the list and the planning of the intended validation exercises.

Chapter 3 Conduct of Validation Exercises – reports the details of each validation exercise in term of preparation and execution.

Chapter 4 Validation Results – reports the summary of exercise results and the related level of confidence.

Chapter 5 Conclusions and Recommendations – provides the conclusions and final recommendations of the whole study.

Chapter 6 References – contains all the applicable and the reference documents that have been used to support the development of this document.

Chapter 7 Annex – contains proposed improvements overview and some results concerning training effects and the quality of simulations.

2. CONTEXT OF THE VALIDATION

2.1. CONCEPT OVERVIEW

The GreAT Concept aims to put forward the environmental-friendly concept of air traffic operation in line with the current mainstream of air transport systems and their strategic plans, focusing on the vision of greening. This concept covers short- and long-haul operations. The short-haul part of the project (MWP4) mainly addresses arrival, departure and surface operations, and the long-haul part aims to optimize en-route operations (MWP3). The latter is covered by the Chinese partners; thus, it is not integrated into the present document, but will be addressed in the Validation Report developed by the Chinese consortium (along with the short-haul use case).

The short-haul operation will consider two aerodromes with different sizes and complexity: medium sized airport and hub airport. The detailed concept related to the short-haul operation could be found in the deliverable D4.1 [3]. Here below, only a short overview of the concept elements being tested and assessed within these validation activities is provided:

- **Airspace design**: Based on a new airspace design enabling continuous climb operations (CCOs) and continuous decent operations (CDOs), arrival operations are sequenced, scheduled, and supported with guidance advisories for the air traffic controllers by DLR’s arrival manager (AMAN) called 4D-CARMA (4D cooperative arrival manager).
- **Optimization of ground operations**: Optimized departure operations are scheduled by the departure manager (DMAN) of DLR called CADEO (Controller Assistance for Departure Optimization). Optimized taxi operations are scheduled by the surface manager (SMAN) of DLR called TRACC (Taxi Routing for Aircraft: Creating and Controlling). Through the development of a 4D capable surface management system the conventional departure management system working with standard taxi times is outdated and will be combined with DMAN. SMAN uses a genetic algorithm to plan and adjust taxi-trajectories in real-time to resolve conflicts between aircraft on the ground, with the aim to reduce holding time after engine startup as well as preventable braking and acceleration actions due to other traffic.
- **MergeStrip system** jointly developed by HungaroControl and Pildo Labs: it supports the air traffic controllers to enable CDOs for aircraft. It was further adapted and enhanced to provide a sequence planning of arrival traffic for air traffic controllers considering a what-if functionality.

These concept ideas are addressed separately and jointly towards the reduction of the environmental impact considering safety and the operational performance.

Qualitative and quantitative assessment for the addressed concepts will be helpful to draw conclusions and recommendations for the future.

2.2. SUMMARY OF THE VALIDATION PLAN

The validation plan [Kling 2021] was prepared in line with European Operational Concept Validation Methodology (E-OCVM) [EUROCONTROL 2010] and the validation strategy set out in SESAR 2020 [Brochard 2019]. This chapter will present the main elements of the validation plan so that to ease the referencing and cross checking between the VALP and VALR.

2.2.1. SUMMARY OF VALIDATION EXERCISES

Two exercises were defined for this validation:

- EXE-001 – DLR - Validation of advanced controller support tools at an airport
- EXE-002 – HC and Pildo Labs - Validation of the new MergeStrip 3.0 functionalities

The tables hereafter summarizes the details of both exercises.

Table 1. EXE-001 Details.

IDENTIFIER	EXE-001 – DLR
TITLE	Validation of advanced controller support tools at an airport
DESCRIPTION	This validation put the focus on a coordinated arrival-/departure flow to show the benefit of such kind of system. The used management tools will assist the controllers in coordinating the in- and outbound traffic of an airport. By performing human in the loop simulation in a TMA environment, the environmental improvements, workload, capacity, and safety will be assessed. In addition, the incoming arrival traffic will be validated with automatic simulation regarding an environmental improvement of surface movements by using a surface manager at the tower/ apron control.
EXPECTED ACHIEVEMENTS	The intended new or improved controller assistant tools will reduce the environmental impact of air traffic without reducing safety or capacity or increasing workload for the controller.
USE CASES	airport pair air-to-air
VALIDATION TECHNIQUE	Real-time Simulation Automatic Simulation
KPA/TA ADDRESSED	Environment, Safety, Capacity, Operational Efficiency, Human Performance
START DATE	2021 Q4
END DATE	2022_Q4
VALIDATION COORDINATOR	DLR
VALIDATION PLATFORM	ATMOS and own developed tools
VALIDATION LOCATION	Braunschweig, Germany
DEPENDENCIES	WP4.2

Table 2. EXE-002 Details.

IDENTIFIER	EXE-002 – HC and Pildo Labs
TITLE	Validation of the new MergeStrip 3.0 functionalities

DESCRIPTION	The aim of this validation exercise is to demonstrate the potential benefits of the new functionalities of MergeStrip 3.0 The tool will be used to help ATCOs better sequencing the arriving traffic to Budapest in APP environment. The exercise will assess the potential of MergeStrip 3.0 new functionalities to decrease CO_2 emissions, support ATCOs in their decision-making process, decrease workload, improve situational awareness, shorten flight paths, maintain airspace capacity, and increase flight efficiency while meeting the relevant safety standards at the same time. (In order to avoid repetition and easier reference, MergeStrip with new functionalities developed under GreAT project is sometimes referred to as MergeStrip 4.0. This, from legal point of view, does not mean any sort of differentiation from MergeStrip 3.0 as mentioned in the Consortium Agreement, Attachment 1: Background included, Party 3: HC, under any circumstances.)
EXPECTED ACHIEVEMENTS	• The new functionalities carry environmental benefits. • The new functionalities meet end user' expectations. • The new functionalities earn end-users' trust and confidence and are working safely.
USE CASES	airport pair
VALIDATION TECHNIQUE	Real-time Simulation
KPA/TA ADDRESSED	Environment, Human Performance, Safety, Capacity, Operational efficiency, Cost effectiveness
START DATE	2021 Q4
END DATE	2023 Q2
VALIDATION COORDINATOR	HC
VALIDATION PLATFORM	MATIAS-BEST, Operations Room
VALIDATION LOCATION	Budapest, Hungary
DEPENDENCIES	WP4.2

2.2.2. SUMMARY OF ADDRESSED KEY PERFORMANCE AREAS

The validation activities performed within MWP6 aims to bring together the various performance areas and validate the developed airspace design and ATC support tools, and thus provide feedback on the overall concept applicability. The main Key Performance Areas (KPAs) for the European Consortium to be addressed are:

- **Environment**: Analyze the improvement of various emissions with a single flight and airspace.
- **Operational Efficiency**: including improvements in flight trajectory efficiency, flight time efficiency, and efficiency of free-route airspace flight services.

- **Cost-effectiveness**: Consider improvements in ATCO productivity
- **Safety**: Evaluate the safety performance of flight operations in case of complexity, controller behavior/responsibility changes, trajectory updates based on upper weather conditions, and mixed airspace operations under free route airspace and operational rules.
- **Human Performance**: Address the acceptability of procedures, roles, and responsibilities; the suitability of technical systems in supporting the human actor; the adequacy of the team structure and team communication, relevant transition factors.
- **Capacity**: Addressing the impact of the new structure of airport terminal airspace for short haul operation on airport capacity in hub and medium airports and their respective TMA.

MWP6 ran jointly with MWP4 and applied the Human Centered Design methodology [DIN ISO 9241 2020] to complement E-OCVM by ensuring that the developed system and airspace design solution meets user requirements.

2.2.3. SUMMARY OF VALIDATION OBJECTIVES AND SUCCESS CRITERIA

To increase the consistency within the GreAT activities, harmonized validation objectives and success criteria have been developed for the European Consortium. The validation objectives are related to the KPAs defined in the previous chapter.

Success criteria will be measured depending on the KPA Category.

Being an **environmental** focused project, the focus of this validation is the reduction of fuel burn, and thus greenhouse gases emissions, first of all CO_2. However, under MWP7, other emission types, e.g., NOx will be examined. Close cooperation between MWP6 and MWP7 is established by Project Partners. The final action plan for the application of the model to be developed under MWP7 will be determined in a later phase, adjusted to the model's development.

Operational efficiency strongly depends on the actual flown distance (or on time): the less an aircraft must fly the less fuel is burnt. This is especially true when holdings and level flights, the most fuel consuming options must be applied in order to respect the relevant safety rules. Overlapping with environmental objectives was avoided by a clear-cut distinction as under operational efficiency, the fuel burnt will be measured.

Capacity: The specific developments under the GreAT project should have no significant impact on capacity. However, any major operational change or development must enable stakeholders to expect at least the same capacity values.

For **human performance**, standardized questionnaires such as Bedford Workload Scale or SASHA-Q will be applied. Answering with the category "acceptable" (or similar) will indicate success based on majority of answers for those objectives. In other cases, it depends on the expert judgement of the feedback in questionnaires & debriefs. If the majority of the ATCOs and runs provide satisfactory or higher ratings, these results indicate success on the objective. Feedback during debriefs will support expert judgement on the results.

Safety is again a crucial objective for the developed greener operations. It is closely interlinked with Human Performance (de-briefings, questionnaires and simulation logs will be used). Duplication of objectives and criteria that might be assigned to both KPAs was avoided emphasizing human performance. These harmonized validation objectives should be followed by all the validation exercises, although differences in the focus can be expected

No development can gain ground unless proves to be **cost-effective**. In this regard, the change in ATCO productivity will be examined, an increase or at least up keeping the current ATCO productivity level is expected.

The following list provides an overview on the generic validation objectives and validation criteria used for validating the developed airspace design and ATC support tools. The exercise-specific objectives and success criteria can be found in the following chapters.

Figure 1. Overview of Common Objectives and Success Criteria.

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	DLR EXE	HC-Pildo EXE
ENVIRONMENT					
ENV–GREAT-01	To assess the reduction of exhaust emissions due to solution	ENV–GREAT–CRT-01-10	The solution results in reduction of exhaust emissions compared to the reference scenario.	X	X
OPERATIONAL EFFICIENCY					
OPE–GREAT-02	To assess the reduction in flown distance per aircraft due to solution	OPE–GREAT–CRT-02-10	The distance flown is reduced compared to reference scenario.	X	X
OPE–GREAT-03	To assess reduction in fuel-burn due to solution	OPE–GREAT–CRT-03-10	The average fuel burn by aircraft is reduced compared to the reference scenario.	X	X
CAPACITY					
CAP–GREAT-04	To assess the solution’s impact on capacity	CAP–GREAT–CRT-04-10	The solution does not reduce capacity.	X	X
HUMAN PERFORMANCE – WORKLOAD					
HUM–GREAT-05	To assess the ATCO’s workload	HUM–GREAT–CRT-05-10	The level of workload is within acceptable limits.	X	X
HUMAN PERFORMANCE – SITUATIONAL AWARENESS					
HUM–GREAT-06	To assess the ATCO’s situational awareness	HUM–GREAT–CRT-06-10	The level of situational awareness is within acceptable limits.	X	X
HUMAN PERFORMANCE – USABILITY					
HUM–GREAT-07	To assess the usability of the system	HUM-GREAT-CRT-07-10	There is no discrepancy between system-provided information and user-required information.	X	X
		HUM-GREAT-CRT-07-20	The ATCO can perform interaction without noticeable problems.	X	X
		HUM–GREAT–CRT-05-30	The alarms and alerts support task performance.	N/A	N/A

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	DLR EXE	HC-Pildo EXE
		HUM–GREAT–CRT-05-40	The look-and-feel of the HMI is acceptable.	X	X
HUMAN PERFORMANCE – TRUST					
HUM–GREAT-08	To assess the ATCO's trust in the system	HUM–GREAT–CRT-08-10	The level of trust is experienced as sufficient by the ATCO.	X	X
SAFETY					
SAF–GREAT-09	To assess the impact on the safety level of the system	SAF–GREAT–CRT-09-10	Procedures and system functions are safe in normal situations.	X	X
		SAF–GREAT–CRT-09-20	Procedures and system functions are safe in abnormal situations.	N/A	X
		SAF–GREAT–CRT-09-30	Procedures and system functions are safe in degraded mode situations.	N/A	N/A
COST EFFECTIVENESS					
COS–GREAT-10	To assess the impact on ATCO productivity	COS-GREAT–CRT-10-10	ATCO productivity is not decreased compared to the reference scenario.	N/A	X

2.2.4. SUMMARY OF VALIDATION USE CASES AND SCENARIOS

Two main use cases are planned in the short-haul operations to test the developed ATC support tools to their full extent:

1) **airport-pair use case**: a short haul flight between an airport pair of a hub airport and medium sized airport;

2) **air-to-air use case**: a hub airport considering an approach, taxi-in, turn-around, taxi-out and departure;

The key variables that the exercises may pick from are the followings:

Maturity of the ATC decision tool	•Reference •New tool
Airspace design	•Reference •New design
Traffic Load	•Medium, in accordance with the context •High, in accordance with the context
Operational mode	•Normal operational mode •Abnormal scenario •Degraded mode

Figure 2. Variables and their levels on solution-level.

2.2.4.1 REFERENCE SCENARIO

Having a reference scenario is considered important to ensure comparability. The baseline for the GreAT concept is the current operations as described in the CONOPS of MWP2. However, the respective exercises differ in their operational environment, thus will not be able to start from the same basis. The details can be found in the description of the respective Exercise.

The baseline scenario for EXE-001 can be either created by performing simulations without additional GreAT tools and without the new airspace structure, or by analysing databases from real air traffic (e.g., OpenSky) and evaluate them according to the selected key performance indicators. Depending on the availability and quality of data, the second option will be used. First analysis shows variations of traffic patterns. For the validation, the data of the last 14 month before the Covid crisis will be assessed to get a statistically representative view, which is coherent with the modelled scenarios.

2.2.4.2 SOLUTION SCENARIO

The project could be seen as a set of solutions and concept ideas collected within MWP2 and MWP4 which may help in reducing the environmental footprint of aviation. Some of these concepts were selected to be tested and evaluated within these validation activities and then they could be seen as different solution scenarios. The following summary represents these solutions:

- **New airspace design** enabling CCOs and CDOs and **its supporting tools** (This will be tested through airport-pair scenario)
- **Optimized taxi operations** (This will be tested through air-to-air scenarios of a hub airport)
- **MergeStrip system** (This will be tested through airport-pair scenario)

All these solutions will be tested separately or jointly within different runs to try to ass the potential benefits from them.

2.2.5. VALIDATION EXERCISE EXE-001 – DLR

The validation exercise EXE-001 will be performed in two iterations. This allow to collect the first set of feedbacks on the tools/ chosen scenarios as well to detect any failure or errors to be fixed before the final run later on. This will help to achieve more reliable results by rectifying errors and adjusting parameters and configurations.

2.2.5.1 VALIDATION EXERCISE DESCRIPTION AND SCOPE

- **REAL-TIME SIMULATION**

The airport selected for the airport-to-airport connection is Munich Franz-Josef-Strauß airport (EDDM). The exercise will focus on the new Extended TMA structure and the associated assistance tools. It will be run on the NARSIM (NLR's Air traffic management Real-time SIMulator) within the Air Traffic Validation Center of DLR in Braunschweig/ Germany. With the new improved controller assistance system 4D-CARMA, it will be possible to separate aircraft regarding their equipage with 3D-FMS and 4D-FMS and to guide 4D-FMS aircraft to the final without time and energy consuming trombone approaches or holding patterns. Instead of the manual guidance with multiple clearances for speeds, altitudes and headings, the AMAN will act proactive with Early Full Clearance Approach (EFCA) advices to avoid unnecessary fuel burn and greenhouse gas emissions.

A cooperative master-slave connection between AMAN and DMAN based on the Fuzzy-rule support system ADCO ensure the optimal balance between inbounds and outbounds on the same or on two dependent runways without resource conflicts. This ADCO controlled cooperation reduces waiting times both on the ground and in the air, thus reducing fuel consumption and gas emissions.

- **AUTOMATIC SIMULATION**

A simulation with the SMAN TRACC will validate the benefit of a surface management system for ground movements at a hub airport in automatic mode. For departing aircraft, the DMAN CADEO will ensure that no waiting times at the runway holding point will occur and each aircraft will climb according to its own optimal profile. The cooperation between DMAN and SMAN allows a conflict-free route planning for each individual aircraft from gate to runway and vice versa with advisory support for controllers to transmit the according aircraft clearances at the right time

For both simulation activities, different scenarios, composed of varying traffic volume with variable arrival departure ratio will be processed. By running it with similar traffic compared to the reference scenario for both arrivals and departures, the intended validation objectives will be analysed by comparing the estimated fuel burn of arrivals as a summation of the complete scenario.

2.2.5.2 VALIDATION SCENARIOS

The design of the experiments has different variables, which describes a single scenario. For each variable one value is selected for a single simulation run. For EXE-001, two solution scenarios are envisaged as per the used validation technique.

REAL-TIME SIMULATION

Human-in-the-loop simulations are performed. These are much more elaborate than automatic simulations, but allow much more detailed conclusions to be drawn about the usability of the proposed systems used their interactions and features.

AUTOMATIC SIMULATION

This simulation works completely in automatic mode without any humans involved. This solution serves to carry out as many simulations as possible in order to be able to make statistically relevant statements at the end. This also allows as many different traffic situations as possible to be covered, demonstrating that the support systems involved can cope with a wide range of scenarios and propose safe solutions. However, automatic simulations only allow very limited statements regarding the usability by human controllers.

It should be noted that the human-in-the-loop (HITL) validation runs will also enable the calibration of the automatic simulations, so that the results of the automatic simulations can be extrapolated within certain limits with the help of the HITL experiments and thus ideally increasing their validity. Due to the effort involved in HITL simulations, it may not be possible to test all support systems in extensive simulation runs.

Both validation scenarios will be conducted for medium and high traffic scenarios and for three different aircraft equipment scenarios (Figure 3).

Use of airspace and automatization of ATC decision tools	• **Reference:** Without tools in existing airspace structure • **Solution I:** AMAN, DMAN, SMAN & ADCO work complete automated with NARSIM • **Solution II:** AMAN supports human controller, DMAN, SMAN & ADCO work automated
Traffic load	• Medium (ARR/h TBD) • High (ARR/h TBD)
Traffic distribution	• 0% A-FMS equipped aircraft scenario • 40% A-FMS equipped aircraft scenario • 75% A-FMS equipped aircraft scenario

Figure 3. Variables and their levels in the trial scenarios for AMAN and airspace design evaluation.

2.2.5.3 VALIDATION OBJECTIVES

In this project, the main focus is set on the analysis of the environmental impact by assessing possible fuel burns savings, which directly impact the CO_2 emissions. In addition to this project goal driven validation topic, other key performance areas (KPA) like safety, capacity, efficiency, and human performance will be analysed. The latter is not a KPA according to ICAO [ICAO 2005], but considered as one by SESAR [Grier 2015].

Therefore, it is considered here. These KPAs are necessary to assess to prove that the positive environmental impact has no negative consequences for these other areas.

Fuel burn will be accessed to determine the greenhouse gas emissions, which are commensurate to the burned fuel. This will be done in a first step with existing in-house tools. In parallel the necessary trajectory information to calculate emissions in a more sophisticated way are sent to WP7 to validate and improve the first step results. Other key performance areas are determined by interviewing the controllers after the trials and by analysing the simulation log files, which include the trajectory associated with thrust settings of each aircraft in detail. Furthermore, if appropriate the corresponding radio communication can be evaluated.

By calculating the flown distances in the TMA, a comparison with a baseline can be made. To allow the comparison of high demand situations, the baseline will be created by analysing pre-Covid traffic from January 2019 to February 2020 extracted from the OpenSky historical database [OpenSky 2023]. Here the focus will be put on the selected airport pair (Munich Franz Josef Strauß airport – Budapest Liszt Ferenc airport).

The detailed list of Validation Objectives addressed in Validation EXE-001 are provided in the following Table.

Table 3. EXE-001 validation objectives, success criteria and how to address description.

Validation Objective	Success criteria	How will it be addressed? (e.g Log Analysis, Questionnaires, Debriefings, Observation)
ENVIRONMENT		
To assess the reduction of CO_2 due to improved management systems	Less fuel burned as average of complete traffic scenario compared to reference scenario	Simulation Log
SAFETY		
To assess the usability and trustworthiness of the new developed tools	The ATCO uses the new tools with confidence	Questionnaires and debriefing
To assess the separation of each aircraft pair	There are no critical separation infringements	Simulation log
Operational EFFICIENCY		
To assess the efficiency of the new system (airspace structure & controller assistance tools)	The flown distance within an observation horizon as average of complete traffic scenario is reduced compared to baseline	Simulation log

Validation Objective	Success criteria	How will it be addressed? (e.g Log Analysis, Questionnaires, Debriefings, Observation)
CAPACITY		
To assess the impact of the advanced controller assistance tools on the capacity	There will be no drop-in capacity	Simulation log
HUMAN PERFORMANCE		
To access the ATCO's workload	The ATCO can handle the traffic without excessive workload over a defined period	ISA (Instantaneous Self-Assessment) & questionnaires

2.2.5.4 VALIDATION METRICS

Depending on the availability of controllers it is intended to have runs with ten different controllers in real-time simulation. Each controller will work one day with the system, depending on his/her availability either from noon one day to noon the other day or from morning to evening at one day. Each controller will be presented different scenarios according to Figure 3.

During the simulation runs, the controllers will be requested to put in a number (usually from 1 to 5, i.e., bored to excessive workload) in a touch monitor, which is connected with the voice communication system. This self-assessment gives an indication of the perceived workload.

After the simulation, questionnaires and interviews with the involved test persons will be conducted to get a deeper view on KPIs like acceptance, confidence, or safety. In addition, validation experts will observe the test persons during the runs.

Data from the simulation log will be assessed and analysed with state-of-the-art tools. This process allows a quantitative analysis of air traffic performance data and therefore, is an important cornerstone of the collection method.

Table 4. Overview data collection method per KPA for EXE-001.

KPA	KPIs	Metric / Indicator	Method / Technique
ENV	**Average fuel burn per flight**	Comparison of average fuel burn between reference and advanced scenarios	Simulation log
EFF	**Flight distance in TMA**	Comparison of flown trajectories regarding flight distances as average summary of complete scenario	Simulation log
CAP	**Number of operations per unit of time**	Number of arrivals and departures per unit of time	Simulation log

KPA	KPIs	Metric / Indicator	Method / Technique
HP	**Workload**	ATCO perceived workload	ISA test
SAF	**Safety performance**	Number of separation minima infringements; perceived level of safety	System logs ; Post-validation questionnaire ; debriefing
	Confidence in using the new tools	Perceived reliability, integrity, and usability of support functions	Post-validation questionnaire; Debriefing

2.2.6. VALIDATION EXERCISE EXE-002 – HC AND PILDO LABS

2.2.6.1 VALIDATION EXERCISE DESCRIPTION AND SCOPE IN RTS

For the medium-sized airport, the simulated environment represents the new Budapest TMA implemented in January, 2020. The exercise thus focuses on the ATC decision support tool: it is planning to develop and demonstrate the extended use of MergeStrip to include additional functionalities (e.g., what-if probing support, precise arrival time estimation and conflict resolution recommendation). The validation will be run on HungaroControl's MATIAS-BEST Real-Time simulator, with 4 ATCOs in two iterations.

The key objective is to investigate whether this support tool can achieve the environmental goals in terms of CO_2 set in the Grant Agreement, under Section 2.2.3, Common Validation Objectives and under Section 2.2.6.4, MergeStrip specific validation objectives of this Validation Plan.

An equally important goal is to see the Human Performance angle of this ATCO decision support tool. To this end, the validation team must check whether MergeStrip can reduce ATCO's cognitive workload and support their situational awareness to be able to make more informed decisions, especially concerning environmental consequences.

This will be achieved by improving the calculation of the Estimated Time of Arrival (ETA). New techniques based on data analysis will improve the accuracy of the ETA estimation, allowing ATCOs to precisely sequence the arrivals at a very early stage and therefore enhancing the use of full CDOs (starting as close as possible to the Top of Descent).

Another main evolution planned for the new MergeStrip is the implementation of the "what-if" functionality. This feature will allow ATCOs to analyse the consequences of any potential action before executing it (e.g., applying speed control / changing target waypoint). The impact on the overall scenario in terms of fuel consumption and CO_2 emissions will be one of the main outputs of the "what-if" analysis.

The final evolution targets the utilization of data analysis to address and propose optimal solutions to potential conflicts. Nowadays, the most used techniques to keep separation and sequence between aircraft arriving to an airport during traffic peak scenarios are based on Standard Arrival Routes vectoring and (in worst cases) the use of holding patterns. These techniques are far from being optimal from the operational and environmental points of view (increased fuel consumption, flight delays, unpredictability etc.). By making use of data analysis techniques, MergeStrip will recommend ATCOs more optimal solutions based on the application of speed control or target waypoint change at an early stage of the approach, allowing to maintain the runway throughput while avoiding non-optimal tactical interventions of ATCOs during descent.

All other KPAs enlisted in the Grant Agreement will be addressed in this validation activity as well, i.e., Operational efficiency, Capacity and Cost-effectiveness. Last, but not least, the new MergeStrip functionalities present a very specific case from Safety point of view, as most of them are based on machine-learning algorithms highly dependent on the built models. Also, the possibility of programming malfunction alerts into MergeStrip is limited, which means that the conventional degraded mode validations are not possible. Therefore, this angle can be only determined once the development of these functionalities is carried out, which is scheduled after the submission of this present Validation Plan. This aspect must be revisited later when the development of these functionalities is in a more mature phase, e.g., in the framework of a workshop. The findings of EASA Concept Paper: First usable guidance for Level 1 machine learning applications will be used. Also, there is an ongoing standardization work in EUROCAE, whose findings are planned to be incorporated. The current understanding of the partners is that safety focus should not be put on degradation of functions but on system reliability and the risk that system is providing misleading information. Simulation Logs are planned to be used but this is conditional to software development under WP4.2

It must be noted that even tough under the environmental KPA, there is only one validation objective, altogether 48 runs were planned in the two iterations for a well-established performance assessment of MergeStrip.

Human Performance and Safety will be assessed from multiple angles, as three new functionalities are developed under WP4.2., and they are intended to bring significant improvements. Also, in the case of degraded mode, it is not yet possible to determine which functionality can be validated hence success criteria concerning degraded modes might be changed accordingly in later stages of the project.

2.2.6.2 VALIDATION EXERCISE DESCRIPTION AND SCOPE IN SHADOW MODE

The last validation activity for Budapest was a shadow-mode validation in the OPS room. The exercise focused on the ATC decision support tool (MergeStrip) with its additional functionalities (e.g., 'What-if' probing support, improved arrival time estimation and conflict resolution recommendation). The validation lasted for 10 days, 29th March- 7th April, 2023.

In the OPS room a test Control Working Position has been set up to properly test the new MergeStrip in a live environment and also to compare it with the current version of the MergeStrip. Four and a half hours per day has been assigned for the validation (between 0945-1130, 1545-1700 and 2030-2200 (UTC) adjusted to the anticipated peaks of LHBP arrivals) and the ATCO roster has been accommodated accordingly. It meant 63 hours of validation altogether, and it must also be mentioned that the application was still left in the OPS room up and running in case someone individually wanted to take a try with it.

Eleven ATCOs received training on the new version and were asked to provide their feedback. Almost all of them has participated in the previous simulations thus had also a basis for comparing not only the two MergeStrips, but also to check whether their improvement recommendations from the simulations have been integrated into the shadow-mode validation version.

The key objective was to investigate whether this support tool can achieve the environmental goals in terms of CO_2 set in the Grant Agreement, under Section 1.1, Common Validation Objectives and under Section 3.1.1.

An equally important goal was to see the Human Performance angle of this ATCO decision support tool. The simulations had already addressed workload and situational awareness, but the shadow mode validation enabled to further address the usability of the new functionalities that the team could not test in the previous validation activities.

Among the new functionalities was the improved calculation of the Estimated Time of Arrival (ETA). New techniques based on data analysis aimed to improve the accuracy of the ETA estimation, allowing ATCOs to precisely sequence the arrivals at a very early stage and therefore enhancing the use of full CDOs (starting as close as possible to the Top of Descent).

The next evolution targeted the utilization of data analysis to address and propose optimal solutions to potential conflicts. Nowadays, the most used techniques to keep separation and sequence between aircraft arriving to an airport during traffic peak scenarios are based on Standard Arrival Routes vectoring and (in worst cases) the use of holding patterns. These techniques are far from being optimal from the operational and environmental points of view (increased fuel consumption, flight delays, unpredictability etc.). By making use of data analysis techniques, MergeStrip aimed to recommend ATCOs more optimal solutions based on the application of speed control or target waypoint change at an early stage of the approach, allowing to maintain the runway throughput while avoiding non-optimal tactical interventions of ATCOs during descent.

Lastly, the 'What-if' functionality was validated with live traffic feed as well, although the tool has already been tested twice in the simulations. This feature allowed ATCOs to analyse the consequences of any potential action before executing it (e.g., applying speed control / changing target waypoint). The impact on the overall scenario in terms of fuel consumption and CO_2 emissions will be one of the main outputs of the "'What-if'" analysis.

The new MergeStrip functionalities present a very specific case from Safety point of view, as the improved ETA is based on machine-learning algorithms highly dependent on the built models. Also, the possibility of programming malfunction alerts into MergeStrip is limited, which means that the conventional degraded mode validations are not possible. This aspect must be revisited later when the development of these functionalities is in a more mature phase, e.g., in the framework of a workshop. The findings of EASA Concept Paper: First usable guidance for Level 1 machine learning applications will be used. Also, there is an ongoing standardization work in EUROCAE, whose findings are planned to be incorporated. The current understanding of the partners is that safety focus should not be put on degradation of functions but on system reliability and the risk that system is providing misleading information.

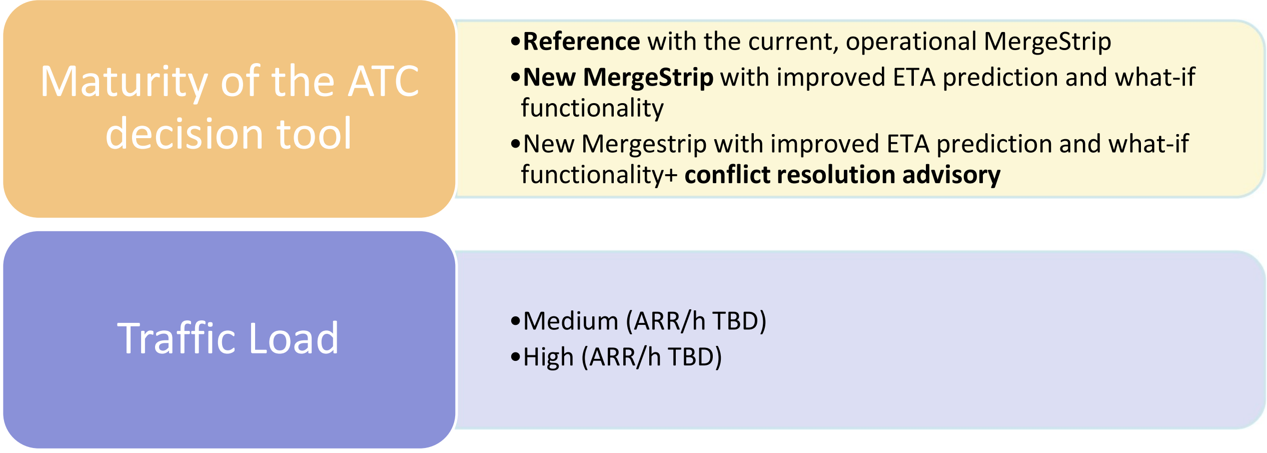
2.2.6.3 VALIDATION SCENARIOS

Validation scenarios are separated in real time simulations (RTS) and shadow mode trials.

2.2.6.3.1 VALIDATION SCENARIOS IN RTS

In the experimental design of the first iteration, there are two variables. The first (Maturity) has three levels, whilst the second (Traffic load) has two levels, as follows.

First iteration:



Maturity of the ATC decision tool

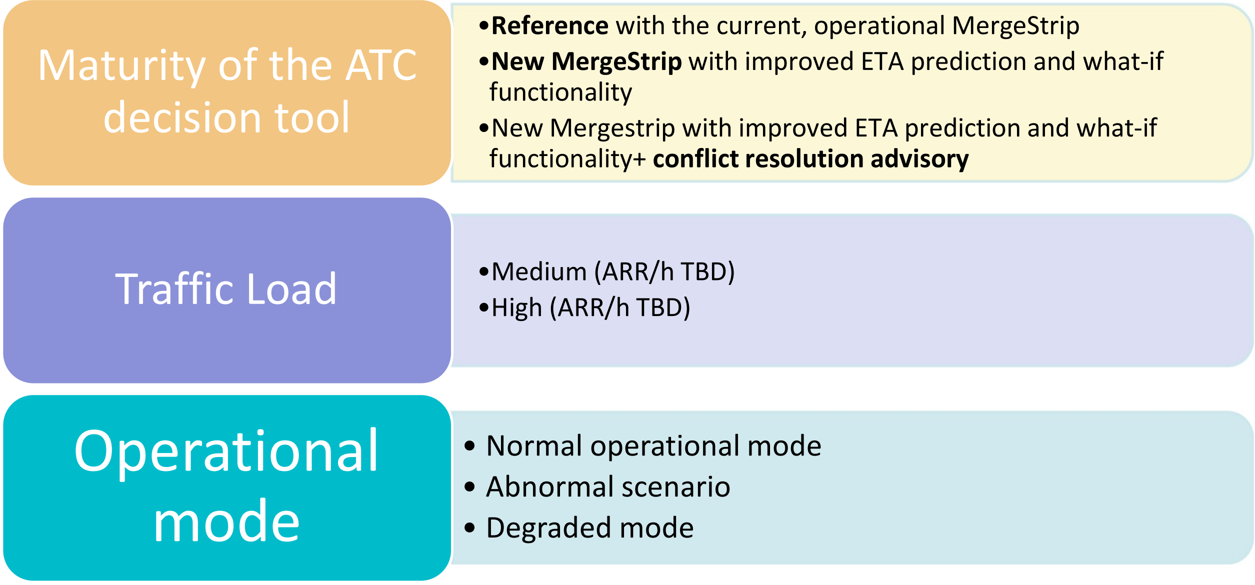
- **Reference** with the current, operational MergeStrip
- **New MergeStrip** with improved ETA prediction and what-if functionality
- New Mergestrip with improved ETA prediction and what-if functionality+ **conflict resolution advisory**

Traffic Load

- Medium (ARR/h TBD)
- High (ARR/h TBD)

Figure 4. Variables and their levels in the simulation.

The second iteration aims to improve the maturity of the solution by covering abnormal scenarios and degraded mode as well. The number of runs with abnormal scenarios and degraded mode are subject to discussion (please refer to Section 4.2.1).



Maturity of the ATC decision tool

- **Reference** with the current, operational MergeStrip
- **New MergeStrip** with improved ETA prediction and what-if functionality
- New Mergestrip with improved ETA prediction and what-if functionality+ **conflict resolution advisory**

Traffic Load

- Medium (ARR/h TBD)
- High (ARR/h TBD)

Operational mode

- Normal operational mode
- Abnormal scenario
- Degraded mode

Figure 5. Variables and their levels in the simulations.

2.2.6.3.2 VALIDATION SCENARIOS IN SHADOW MODE

Not applicable.

2.2.6.4 VALIDATION OBJECTIVES

Validation objectives are separated in real time simulations (RTS) and shadow mode trials.

2.2.6.4.1 VALIDATION OBJECTIVES IN RTS

Table 5. EXE-002 validation objectives, success criteria and how to address description.

Validation Objective	Success criteria	How will it be addressed? (e.g. Log Analysis, Questionnaires, Debriefings, Observation)
ENVIRONMENT		
To assess the reduction of exhaust emissions due to solution	Less CO_2 emitted compared to reference scenario	DailyFuel
OPERATIONAL EFFICIENCY		
To assess the reduction in flown distance per aircraft due solution	The flown distance is reduced compared to the reference	Simulation log
To assess the reduction in fuel burnt per aircraft due solution	The fuel burnt is reduced compared to reference.	Daily Fuel
CAPACITY		
To assess the solution's impact on capacity	The solution does not reduce capacity.	Simulation log
HUMAN PERFORMANCE – WORKLOAD		
To assess the ATCO's workload.	MergeStrip in general reduces the ATCO workload. New functionalities do not increase ATCO workload	Questionnaires (Bedford Workload Scale), Observation, Debriefing
	The what-if function reduces the cognitive workload by supporting the ATCO to find the most optimal solution.	Questionnaires (Bedford Workload Scale), Observation, Debriefing
	The conflict resolution advisory reduces ATCO workload by presenting the most optimal resolution(s).	Questionnaires (Bedford Workload Scale), Observation, Debriefing

Validation Objective	Success criteria	How will it be addressed? (e.g. Log Analysis, Questionnaires, Debriefings, Observation)
HUMAN PERFORMANCE – SITUATIONAL AWARENESS		
To assess the ATCO’s situational awareness.	The new ETA prediction improves the ATCO’s situational awareness *by calculating with more accurate data*	Questionnaires (SASHA-Q), Debriefing
	The what-if function enables ATCO’s to make decisions more efficiently.	Questionnaires (SASHA-Q), Debriefing
HUMAN PERFORMANCE – USABILITY		
To assess the usability of the system.	The improved ETA prediction supports more efficient task performance (arrival sequencing).	Questionnaires (Tailor-made), Observation, Debriefing
	The conflict resolution advisory supports efficient task performance by avoiding non-optimal tactical intervention (i.e., vectoring, holding)	Questionnaires (Tailor-made), Observation, Debriefing, Log Analysis (ATCO inputs)
	Number and/or severity of errors in the solution is within tolerable limits.	Questionnaires (Tailor-made), Observation, Debriefing
	The what-if functionality is easy to interact with.	Questionnaires (Tailor-made), Observation, Debriefing
	The conflict resolution advisory function is easy to interact with.	Questionnaires (Tailor-made), Observation, Debriefing
	The look-and-feel of the HMI is acceptable for the ATCOs.	Questionnaires (Tailor-made), Debriefing
HUMAN PERFORMANCE – TRUST		

Validation Objective	Success criteria	How will it be addressed? (e.g. Log Analysis, Questionnaires, Debriefings, Observation)
To assess the ATCO's trust in the system.	ATCOs trust in the accuracy of the new ETA prediction.	Questionnaires (SATI), Debriefing
	The conflict resolution advisory provided by the system is perceived sensible by the ATCOs. // The conflict resolution advisory provided by the system fits the ATCO's expectations.	Questionnaires (SATI), Debriefing, Observation
SAFETY		
To assess safety of the logic behind system functions in normal situations	According to ATCOs the punctuality of "ETA prediction function" was adequate for safe service provision	Questionnaires, Debriefing, Simulation logs (TBD)
	According to ATCOs the predictions of the "what if function" was adequate for safe service provision	Questionnaires, Debriefing, Simulation logs (TBD)
	According to ATCOs the logic behind conflict resolution advisory was reasonable and adequate for safe service provision (HF-TRUST)	Questionnaires, Debriefing
To assess safety of system functions in normal situations	The working of "ETA prediction function" was appropriate	Questionnaires, Debriefing, Simulation logs (TBD)
	The working of "what if function" was appropriate	Questionnaires, Debriefing, Simulation logs (TBD)
	The working of conflict resolution advisory was appropriate	Questionnaires, Debriefing, Simulation logs (TBD)

Validation Objective	Success criteria	How will it be addressed? (e.g. Log Analysis, Questionnaires, Debriefings, Observation)
	The number of separation minima infringements is not higher	Questionnaires, Debriefing, Simulation logs (TBD)
To assess safety of system functions in abnormal situations	The working of "ETA prediction function" was appropriate	Questionnaires, Debriefing
	The working of "what if function" was appropriate	Questionnaires, Debriefing
	The working of conflict resolution advisory was appropriate	Questionnaires, Debriefing
To assess safety of degraded modes of system functions.	The working of fail-safe operation of "ETA prediction function" is appropriate in case of total/partial loss or corruption of function.	Questionnaires, Debriefing
	The working of fail-safe operation of "what if function" is appropriate in case of total/partial loss or corruption of function.	Questionnaires, Debriefing
	The working of fail-safe operation of conflict resolution advisory is appropriate in case of total/partial loss or corruption of function.	Questionnaires, Debriefing
	The alert in case of degradation of "ETA prediction function" was useful.	Questionnaires, Debriefing
	The alert in case of degradation of "what if function" was useful.	Questionnaires, Debriefing

Validation Objective	Success criteria	How will it be addressed? (e.g. Log Analysis, Questionnaires, Debriefings, Observation)
	The alert in case of degradation of conflict resolution advisory was useful.	Questionnaires, Debriefing
COST-EFFECTIVENESS		
To assess the impact on ATCO productivity	ATCO productivity is not decreased compared to the reference scenario.	Questionnaires

2.2.6.4.2 VALIDATION OBJECTIVES - SHADOW MODE

Validation metrics were the same as in RTS, however, situational awareness, workload and ATCO productivity were not assessed due to the nature of shadow mode validation.

2.2.6.5 VALIDATION METRICS

Validation metrics are separated in real time simulations (RTS) and shadow mode trials.

2.2.6.5.1 VALIDATION METRICS IN RTS

The first iteration of the simulation will be conducted with Real-Time Simulation, and will involve two ATCO pairs and 12 runs/ATCO pairs, thus 6 days in total (or three days if the scenarios are run in parallel). In order to mitigate the learning-effect (seeing the same traffic scenario six times), two very similar versions will be developed for both the medium and high traffic scenarios.

Validation will focus on the safety, human performance and environmental aspects related to the advanced version of MergeStrip in the Budapest TMA environment. Several aspects contributing to effective human performance will be assessed (e.g., situational awareness, workload, user interface and technical systems).

Table 6. Overview data collecting methods per KPA for EXE-002.

KPA	KPIs	Metric / Indicator	Method / Technique
ENV	**Actual Average CO_2 Emission per flight**	Variance in fuel burn as a precursor of exhaust emissions: direct link between fuel burnt and the amount of CO_2 produced (≈3.15 times the mass of fuel burnt)	System log (Daily Fuel web application)
OPEFF	**Actual average Fuel consumption per flight**	Variance in fuel burn as a precursor of exhaust emissions:	System log (Daily Fuel web application)

KPA	KPIs	Metric / Indicator	Method / Technique
OPEFF	**Flight times TMA**	TMA ARR time (Average of the distribution of actual TMA arrival) durations	System log (Daily Fuel web application)
CAP	**Number of arrivals per unit of time**	Number of arrivals per unit of time (hour)	System log (DailyFuel web application)
HP	**Situation Awareness**	ATCO Situational awareness	Post-run questionnaire (SASHA-Q) Debriefing
	Workload	ATCO Cognitive Workload	Over the shoulder observations Post-run questionnaire (Bedford) Debriefing
	Performance of the technical system (i.e., usability, trust)	Effectiveness (Success rate/Errors) Efficiency (time taken to complete a task) User satisfaction (HMI's look and feel) Trust in the performance of the system (e.g., ETA calculation accuracy, reliability of the advisory)	System logs Post-validation questionnaire Over the shoulder observations Debriefing (+ to – adjectives to describe the e.g., concept or HMI)
SAF	**Safety performance**	Number of Separation Minima Infringements Perceived level of Safety	System logs Post-validation questionnaire Debriefing
	Performance of the technical system	Reliability of the functions Integrity of the functions Usability of the functions in normal/abnormal situations Usability of the functions in degraded modes	System logs Post-validation questionnaire Debriefing
CEF2	**ATCO productivity**	Productivity	Post-run questionnaire results will be integrated into the following formula: Increase in productivity (%) = (1/ (1-0.75*workload reduction/2) -1) x100

2.2.6.5.2 VALIDATION METRICS IN SHADOW MODE

Validation metrics were the same as in RTS, however, Situational awareness, Workload and ATCO productivity were not assessed due to the nature of Shadow mode validation.

2.2.7. SUMMARY OF ASSUMPTIONS

The validation assumptions are summarized in the following Table. In the shadow mode validation, only the first assumption was of relevance.

Table 7. Table of Common Validation Assumptions.

Id.	Title	Type of Assumption	Description	Justification	Flight Phase	KPA Impacted	Source	Impact on Assessment
ASM-GreAT-VALP-ALL.01	Training and competencies	Human Performance	All Controllers have appropriate training and competencies.	In order to validate the GreAT concept with the new tools, it is important that the controllers are familiar with the operating environment and tools. It concerns both DLR's simulation environment, and the new software provided by Pildo Labs.	TWRTMA	HP	Expert opinion	High
ASM-GreAT-VALP-ALL.02	Gate-to-gate	Simulator presentation	Those parts of the flight legs that are not examined, will be considered as constant.	Short haul cannot be simulated gate-to-gate, therefore backup solution is needed.	ENR TMA	ENV OP EFF	Runs Post-run calculation System logs (e.g., DailyFuel)	Low
ASM-GreAT-VALP-ALL.03	CO_2 measurement accuracy	Simulator presentation	Reliability of quantitative indicators from simulator	Aircraft performance in simulator may differ from that of real operation	APP	ENV	Runs Post-run calculation System logs (e.g., DailyFuel)	High

2.2.8. CHOICE OF METHODS AND TECHNIQUES

2.2.8.1 EXPERIMENTAL DESIGN

Experimental design for real time simulations (RTS) and shadow mode trials.

2.2.8.1.1 EXPERIMENTAL DESIGN IN RTS

The validation was conducted in the form of real-time human-in-the-loop simulations. Participants conducted several simulation runs with varying traffic distributions (different percentage of 4D FMS). This allows testing the impact of independent variables (e.g. traffic volumes, complexity) on the dependent variables (e.g. mental workload, situational awareness).

Two iterations are conducted per validation Exercise. Feedback from the first iteration was used to modify the system in time for the second iteration.

2.2.8.1.2 EXPERIMENTAL DESIGN IN SHADOW MODE

The shadow-mode validation activity was performed in the OPS room. A separate CWP has been dedicated to test the new MergeStrip, and since the current version was also there, the assigned participant could directly compare the two versions. This means that the Planner Controller was not distracted by the test.

The validation lasted for 10 days, 29th March- 7th April, 2023. 4,5 hours per day has been assigned for the validation and the ATCO roster has been accommodated accordingly.

11 Approach ATCOs received training on the new version and were asked to provide their feedback. Out of 11, 9 ATCOs completed the final questionnaire on Survey Monkey.

2.2.8.2 OBJECTIVE DATA

Objective data for real time simulations (RTS) and shadow mode trials.

2.2.8.2.1 OBJECTIVE DATA IN RTS

The log files of the simulation runs will be used to extract information and data needed to calculate the validation metrics and indicators. They will be mainly used for measuring environmental impact (DailyFuel, CO_2), capacity and safety.

2.2.8.2.2 OBJECTIVE DATA IN SHADOW MODE

As mentioned in D6.2 Verification Plan, the objective data for this validation can be obtained from the DailyFuel application. Daily-Fuel is a web-based performance reporting service using ADS-B data to i) monitor the level of implementation of Continuous Descent Operations, ii) establish fuel consumption baseline on which any improvement could be measurable and iii) report other TMA operational KPIs. DailyFuel is separate or independent from MergeStrip. It processes ADS-B and Mode-S EHS data in binary format. The hourly data files are generated by the PildoBox installed in Budapest for this purpose. As DailyFuel uses real binary data as input, it cannot be used in the simulator environment. Environmental measurements of the What-if functionality will be carried out together with the ML-based functionalities, as the "AI-based sequencing and speed control advisory" can be considered as an extension of the What-if functionality.

2.2.8.3 SUBJECTIVE DATA

Subjective data for real time simulations (RTS) and shadow mode trials.

2.2.8.3.1 SUBJECTIVE DATA IN RTS

Questionnaires were either administered post-run or post-exercise, see Section 3.1.5 for the schedule of the validations. The following standard questionnaires were used:

- The **Instantaneous Self-Assessment (ISA)** was used to obtain a mental workload rating every 5 minutes during the simulation runs [Tattersall 1996]. It consists of a single item that is rated on a 5-point Likert scale ranging from 1 (under-utilised) to 5 (excessive). A mid-level rating of mental workload is desirable while more extreme ratings indicate over- or underload.
- Two questionnaires from EUROCONTROL's "Solution for Human-Automation Partnerships in European ATM" (SHAPE) [Dehn 2008a] were used.
    - The **Situation Awareness for SHAPE (SASHA)** consists of six items measuring situation awareness on a 7-point Likert scale from 0 (never) to 6 (always) [Dehn 2008b]. By inverting the ratings of items 2, 3, 5 and 6 and calculating the mean of all six item ratings, the overall SASHA score is obtained. A higher score represents higher situation awareness and is thus desirable. SASHA was administered post-run.
    - The **SHAPE Automation Trust Index (SATI)** is used to assess the level of trust in a system [EUROCONTROL 2012]. It comprises six items that are rated on a 7-point Likert scale from 0 (never) to 6 (always). An overall score is obtained by calculating the mean of the six items. Each item can also be interpreted individually and represents a different dimension of trust, namely utility, reliability, accuracy, understanding, robustness and confidence. SATI was administered post-exercise.
- The **NASA Task Load Index (TLX)** measures workload on the six subscales mental demands, physical demands, temporal demands, own performance, effort and frustration [Hart 1988]. For this validation, the subscale physical demand was omitted as no physical demand was expected for the task. The subscales were presented in the form of slider bars with 21 gradations each, ranging from 0 to 100 in steps of 5. Raw TLX ratings were used, i.e. the subscales were not weighted. According to Hart, this is a common practice and does not reduce sensitivity [Hart 2006]. An overall raw TLX score was computed by calculating the mean of the five subscale ratings. The NASA TLX was administered post-run.
- The **System Usability Scale (SUS)** [Brooke 1996] was administered to assess usability. The SUS consists out of ten items that are rated on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The individual item ratings have no meaning on their own and are used to calculate a total SUS score ranging from 0 to 100. This is done by calculating item contributions and multiplying their sum by 2.5. For items 1, 3, 5, 7 and 9, the item contributions are the item ratings minus one. For items 2, 4, 6 and 8, the item contributions are 5 minus the item rating. A total SUS score of 0 represents the worst possible usability and a total SUS score of 100 the best possible usability. The SUS was administered post-exercise.

In addition to the standard questionnaires, participants received tailored questionnaires:

- Participants were asked to evaluate their experience in the **simulation** ("I feel well acquainted with the simulation" and "I felt immersed in the simulation") on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) post-training and post-run.
- A final **tailored questionnaire** was administered post-exercise. It contained statements about the technical and procedural features of the system. Participants

rated most statements on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The tailored questionnaire was slightly modified for the second iteration.

The questionnaire data were analysed in a non-parametric and descriptive manner using IBM SPSS Statistics Version 26 [IBM 2019].

- **Over the shoulder observations.** During the sessions, the activities of actors will be observed in order to collect insights about their performance, strategies they use to perform the task and difficulties they experience. In order to better understand the reasoning and the way that provided information is used, operators might be asked to "think-aloud" while performing their tasks.
- **Debriefing.** Semi-structured debriefings will be performed at the end of each validation day, or maybe even briefly before/after lunch. The difficulties on the exercise will be discussed among all the participants (operational, validation and technical staff).

2.2.8.3.2 SUBJECTIVE DATA IN SHADOW MODE

Questionnaires:

Two questionnaires has been designed. A simple paper-based survey with 3 questions has been provided to the participants to write down their instant impressions after the trials. These three questions were the followings:

Which functionalities worked well? In which traffic situations were those useful?

Which functionalities worked not working/worked unexpectedly?

Do you have any further improvement ideas?

The final questionnaire was developed to address the usability related success criteria.

Over the shoulder observations:

During the sessions, the interaction of the participants with the new MergeStrip has been observed in order to collect insights about their performance, strategies they use to perform the task and difficulties they experience. In order to better understand the reasoning and the way that provided information is used, operators might be asked to "think-aloud" while performing their tasks.

Debriefing:

After the validation activity a workshop has been organised to review the results of the final questionnaire and to discuss the participants' impressions about the shadow-mode trial.

2.2.8.3.3 ANALYSIS METHODS

The analysis will be based on the collected qualitative and quantitative data. The RTS scenarios based on the independent variables will be compared in order to see the impact on the dependent variables (Human Performance, Safety, Environment) by applying descriptive statistics and models.

Generated data (system logs) will be used as an input for DailyFuel. The application will provide information on the following Key Performance Indicators:

- Average fuel burnt per flight

- Average CO_2 emissions per flight
- Average flight efficiency per flight
- Number of arrivals per unit of time

3. CONDUCT OF VALIDATION EXERCISES

3.1. VALIDATION EXERCISE EXE-001 – DLR

For the evaluation of the controller decision support tools and the new airspace design, two kinds of trials were performed, which provide different forms of results. In the complete automatic simulation without human interactions, AMAN, DMAN and SMAN were used in an ideal world automatic mode where the aircraft fly the AMAN calculated 4D trajectories as precisely as the NARSIM simulator allows. This presents a perfect world scenario with no significant influence of weather and the engagement of very well-trained air traffic controllers and pilots who master their systems perfectly. The second trial includes humans in the loop, using actively the AMAN support functionalities to guide the standard and manually guided aircraft on the trombone patterns and integrate them into the stream of EFCA aircraft. Both trials are aiming to assess different set of metrics which are complimentary to evaluate the project concept and ideas.

3.1.1. PLATFORM USED IN THE SIMULATION

- **REAL-TIME SIMULATION**

The validation was performed at the Air Traffic Management and Operation Simulator (ATMOS) within the Air Traffic Validation Center of DLR in Braunschweig/Germany. As air traffic generator the generic software NARSIM (NLR’s Air traffic management Real-time SIMulator) was used. For the simulations one controller working position was configurated and included. Besides the simulation traffic handled by air traffic controller and pseudo pilots, automation of additional traffic was performed, too. The air traffic controller was in touch with one to two pseudo-pilots (depending on amount of traffic). The exercise focused on the new Extended TMA structure and the associated assistance tools at the airport of Munich (EDDM).

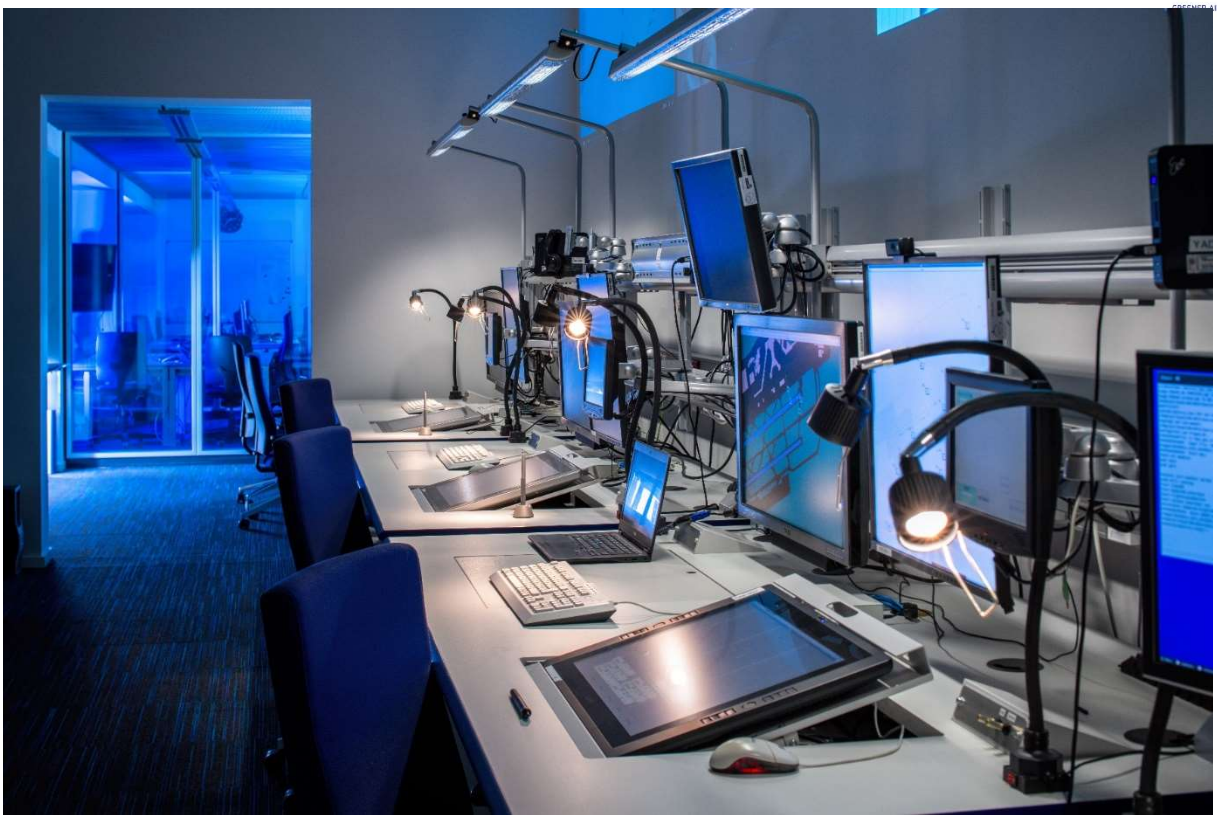


Figure 6. ATMOS facility within the Air Traffic Validation Center in Braunschweig.

AUTOMATIC SIMULATION

The simulation was conducted using the NLR ATC Research Simulator configured for Hamburg airport (EDDH/HAM), the simulation runs were conducted in an automatic mode, without humans in the loop. For the result analysis, it was assumed that the trajectories planned by SMAN based on flight plan data are followed by the pilots, and that no delays or non-conformant behaviour that could cause a re-planning of the trajectories is present.

3.1.2. OPERATIONAL ENVIRONMENT SIMULATED AND NEW TOOLS

The operational environment simulated for the real-time simulation of Munich Airport "Franz-Josef Strauß" is described in Chapter 3.1.2.1 and 3.1.2.2. In addition, the operational environment simulated for the automatic trial is the Hamburg Airport aiming to compare the fuel efficiency of regular taxi trajectories with the optimized conflict-free trajectories generated by the SMAN.

3.1.2.1 OPERATIONAL ENVIRONMENT SIMULATED

The new airspace structure proposed within GreAT is designed for the Munich airport (EDDM) topography with two parallel runways with an offset of 1500 meters. The length of the runways are 4000 meters in both cases, so there are no limitations regarding aircraft types or weather restrictions. The distance between the runways is 2300 meters and therefore they can be used completely independently.

The scenarios used in this Exercise were based on a medium load traffic. The traffic volume equates to around 2/3 of maximum traffic at Munich airport with a reduced number of departures. During trials, departures took place, but do not have to be touched. They are guided in automatic mode by traffic simulator. The traffic mixes (types of aircraft and their frequency of occurrence) for both scenarios orientates on typical EDDM traffic situation from year (2022). Each scenario lasts around 45 minutes.

The GreAT airspace and therefore the planning horizon of the AMAN includes an area with a radius of 150 Nautical miles around the airport and covers the traditional TMA and parts of the adjacent sectors. The runways are used in mixed mode, so that arrivals and departures are executed on both at the same time. Within the GreAT, direct approaches for arrivals to downwind area are possible. The 4D-FMS equipped aircrafts are separated from the standard ones as they proceed directly to LMP perform their optimized arrival profile while the non-equipped A/C use the trombone pattern with downwind, base and final. The benefits of the airspace design results firstly from the possibilities of aircraft to fly a "direct" approach without pressured to use a path stretching area like a trombone or fan pattern, and secondly from the early full clearance approach procedure, which allows pilots to program an aircraft optimal approach procedure like an CDA into the aircraft's advanced FMS without the risk to be interrupted in the approach phase before reaching the Late Merging Point (LMP) on final. Since the share of aircraft with an Advanced FMS currently still varies widely among the individual airlines, the share of existing A-FMSs is used as the scenario difference. In the trial, three shares of A-FMS equipped aircraft are considered: 60%, 30% and 80%. The aircraft equipped with an Advanced FMS get a negotiated target time for the LMP, and conduct an EFCA. In accordance, the non-equipped aircrafts are guides manually by the controllers and use the trombones.

In the same running time of the scenarios, the DMAN-SMAN connection in cooperation with the AMAN and with support of the ADCO must schedule the same number of departures.

3.1.2.2 ATCO SUPPORTING TOOLS

Some supporting and visualizing tools are designed within GreAT. These new tools are necessary as the new airspace structure and its assigned procedures cannot be handled with conventional tools. The exhaustive description could be found in the deliverable D4.1 [Temme 2021]. Here only a short description of each tool/ feature is provided.

- **GHOSTING**

Ghosts represent the theoretical position of 4D-FMS equipped A/C on final. They are only displayed on final and extended centerline. The Ghost position is calculated based on negotiated target time at LMP and not on real 4D-FMS A/C speeds and track miles to LMP. One Ghost represents one 4D-FMS equipped A/C and disappears just before reaching LMP.

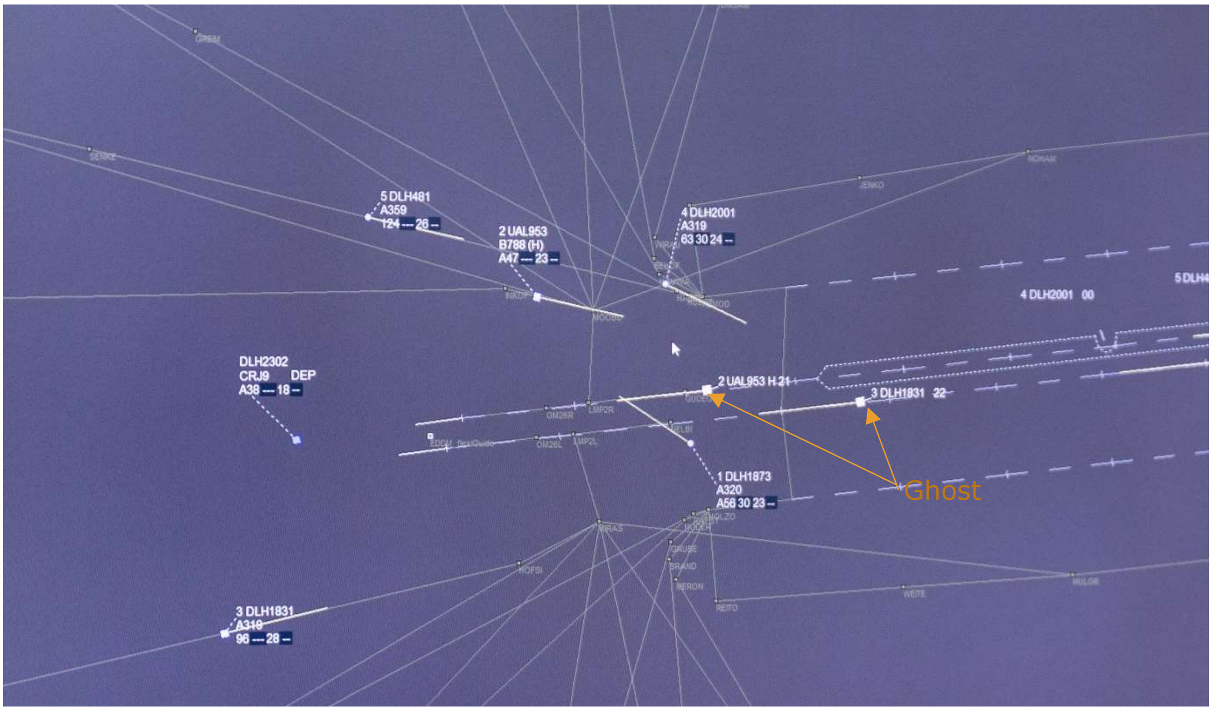


Figure 7. Ghost feature on the radar display.

TARGETWINDOW

TargetWindows ("Targets") represent the optimal position of standard A/C on final. Targets are only displayed on final and extended centreline. The Target position is calculated based on AMAN's 4D trajectories and are conflict free with 4D-FMS equipped A/C. Target Windows position are customized for individual A/C, as the position calculation considers weight classes and speed profiles. The layout of this window is showed in Figure 8. The dotted lined areas around the TargetWindow represent areas with safe separation regarding actual flight planning. ATCOs should guide the standard A/C so that they always meet the TargetWindows as accurately as possible. Targets disappear after an A/C reaches it.

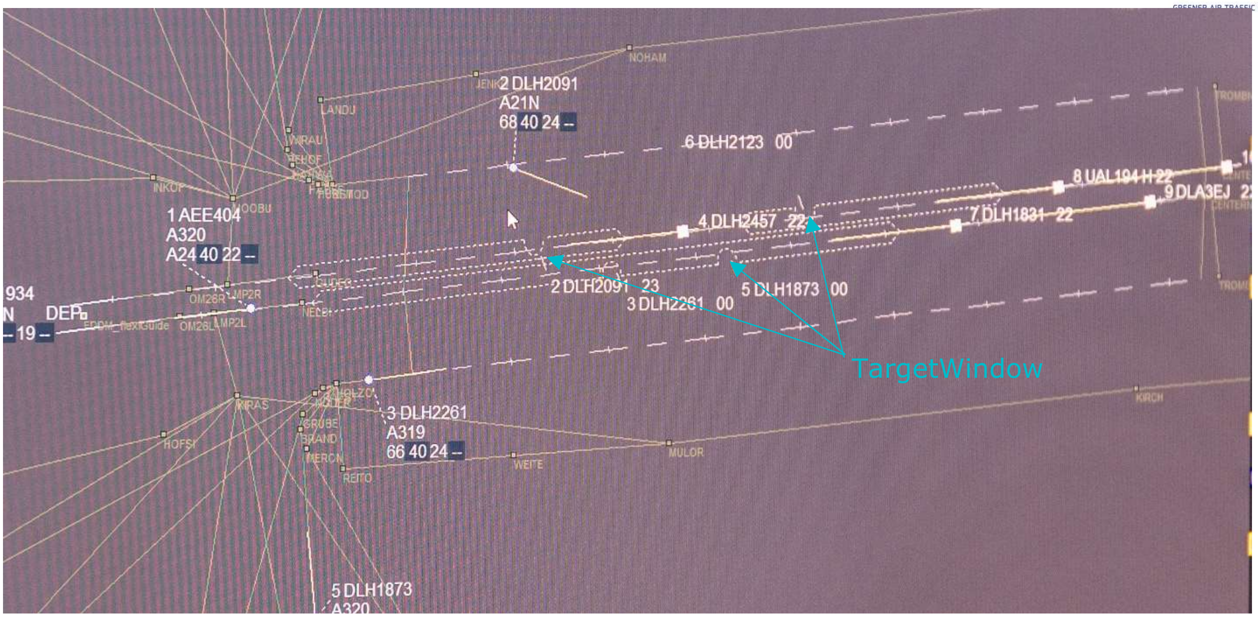


Figure 8. TargetWindow layout with the dotted line indicating the safe area and the semicircle indicating the optimal position.

FINAL DISTANCE INDICATOR

This is an additional support window at bottom of radar screen (Figure 9). It shows actual separation between all aircrafts on Final. It considers aircraft, Ghosts and Target Windows. An aircraft is only viewable, if the aircraft or representative are flying on Final or Centreline.

Located in separate windows for each centerline, the aircraft that are currently on final approach are represented by defined symbols with call signs. In addition, the current distances between the aircraft are displayed in Nautical Miles. In this way, the alphanumeric display enables the controller not only to monitor the current distances, but also to immediately detect any changes in their tendency and to intervene with guidance in the event of imminent separation violations. In addition to actual aircraft, labels for Ghosts (squares) and TargetWindows positions (semicircles) are also displayed, allowing approach controllers to estimate how large the separation will be after turning over the Base Legs or LMP and before reaching the final. Different colours for the weight classes of aircraft allow a more precise differentiation. Green symbols indicating heavy aircraft, yellow ones mediums and small aircraft are white.

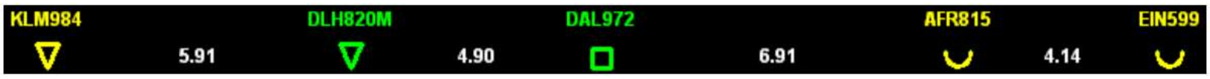


Figure 9. The Centerline Separation Visualization Tool. The symbols mark the position of aircraft (triangle), ghosts (square) and TargetWindows (semicircle). The label colors represent the aircraft weight class (yellow: medium; green: heavy) and the white numbers between the labels indicate the current separation between them.

3.1.3. ROLES & RESPONSIBILITIES IN THE EXERCISE

The following roles are performed during the real-time simulation:

- **ATCOs**: Act as Executive Controller during the Validation runs. Provide feedback about perceived workload and fatigue via the ISA and FISA-system and on questionnaires during and after the runs as well as during the debriefings [Hamann 2020] [Hamann 2022]. In addition to the introduced ATC radar and supporting tools, ISA measure for mental workload is integrated at the CWP on a second touchscreen.

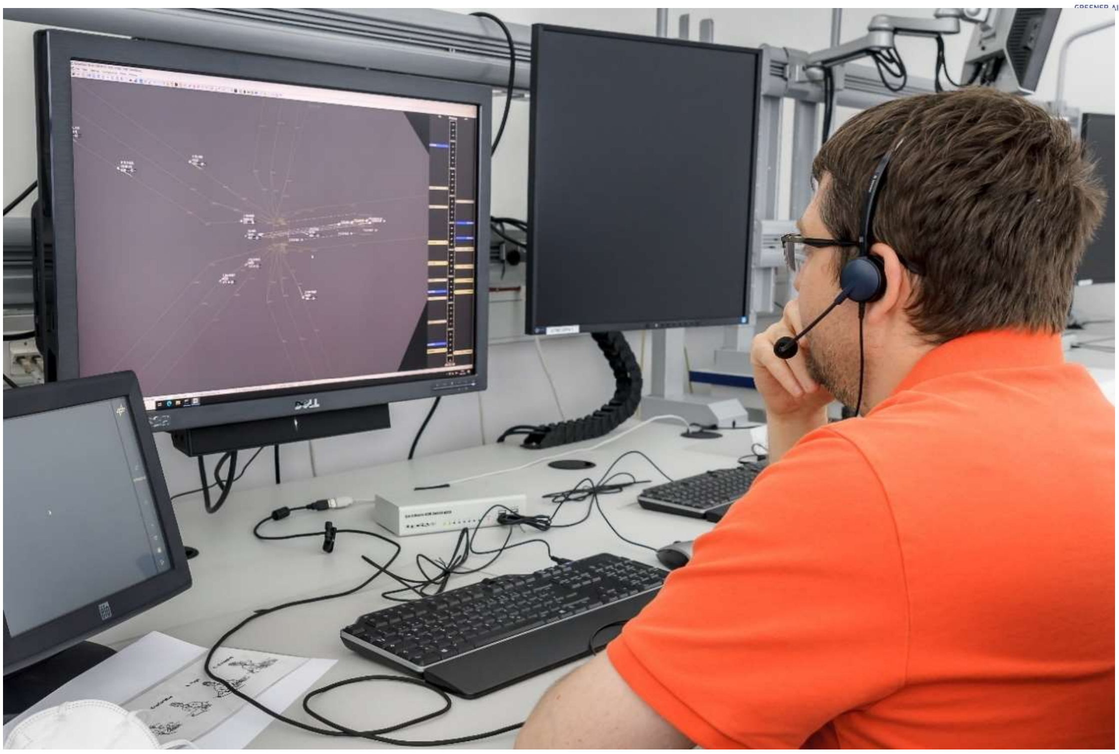
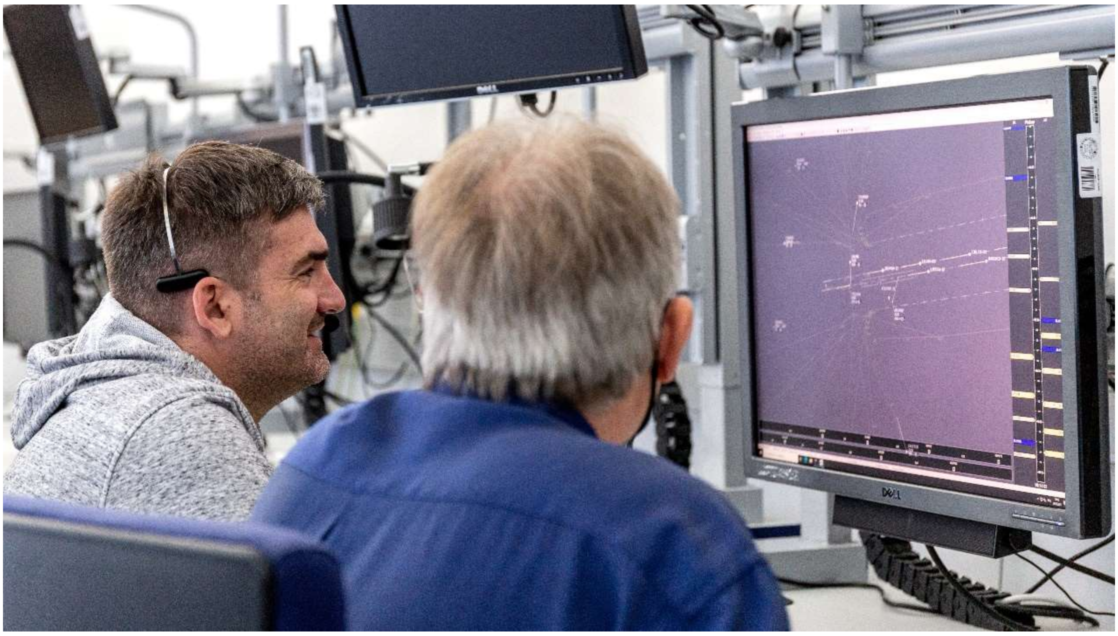


Figure 10. ATCO working on the CWP during a validation exercise

- **Observer/ Expert:** During the sessions, the activities of ATCOs will be observed by an expert in order to collect insights about their performance, strategies they use to perform the task and difficulties they experience. The expert role is also to answer the ATCO questions related to the used tools or concept.



- **Pseudo Pilots**: Provide radio communication to ATCOs during all runs and change aircraft trajectories on dedicated interfaces accordingly. Normally one pseudo pilot controls up to four aircraft depending on the traffic situation and flow. A pseudo pilot is provided with a slightly different display compared to the ATC. The pseudo pilot display basically consists of three parts (Figure 11)

- **Stripview**: Listing all flights radioing on pseudo pilots' frequency.
- **Workspace**: Displaying flight strips of flights under control of the pseudo pilot. Flight strips included aircraft's performance data, such an indicated airspeed, heading, flight level / altitude, arrival route and further more.
- **Radar screen**: Providing an overview of the actual traffic picture within the airspace

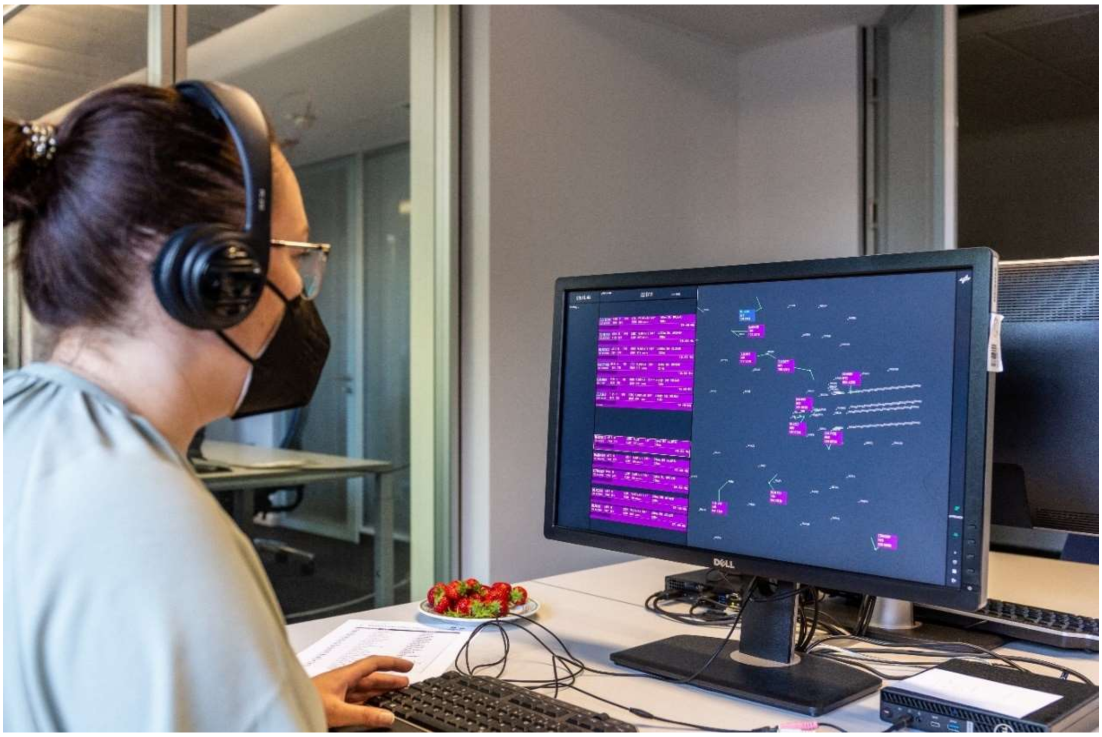


Figure 11. Pseudo-Pilots during a validation exercise

- **Validation Experts**: Preparing and conducting data gathering, and analysing during and after the runs and contributing to the exercise report.

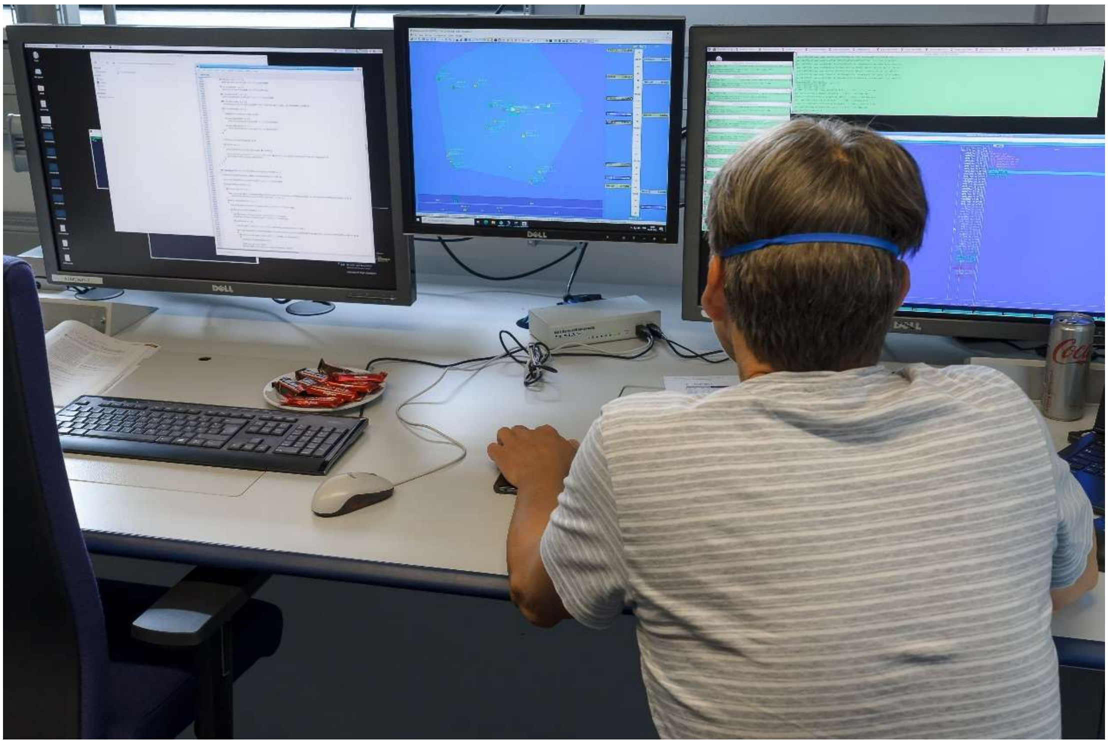


Figure 12. Validation Expert gathering the data during the validation exercise

- **Technical Experts:** Preparing the platforms and scenario files, ensuring proper operation of the platform and all systems necessary for the simulation, supporting the gathering and analysis of quantitative data.

Figure 13. Technical expert ensuring proper operation of the platform and all systems during the different simulation runs

- **Validation Lead:** Ensure timely and coordinated conduction of all runs, coordinate preparation and analysis of validations.

The automatic simulation follows the same schema but with no involvement of controllers or pseudo pilots.

3.1.4. CONTROLLERS' BACKGROUND AND ROLES

During the trials, ATCOs were responsible for standard arrivals (3D FMS equipped or non-equipped). The ATCO`s area of responsibility starts at around 20NM before the downwind and ends after LMP on final around 5 NM before threshold. The ATCO role is the Director (Feeder), and the Executive (Pick-up) is simulated by the traffic simulator and the pseudo pilots. ATCOs were responsible for both independent runways and they guided pseudo-pilots via radio communication.

3.1.4.1 SESSION 1 (MAY 2022)

The sample comprised five male ATCOs from Hungary aged between 32 and 42 years (*M*=39.00, *SD*=4.12). Participants provided written informed consent and received monetary compensation. Their work experience as ATCOs ranged from 3 to 13 years (*M*=10, *SD*=4.12). Four of the participants stated that they have not used functions similar to the ones used in the simulations before, while one participant stated that he uses a system that is similar to the final distance indicator in operation.

3.1.4.2 SESSION 2 (SEPTEMBER 2022)

Five male controllers from Hungary participated in the second iteration of the validation. They were aged between 28 and 44 years with a mean age of *M*=36.80 years (*SD*=8.12). Their average work experience as controllers was *M*=10.00 years (*SD*=7.18), ranging from 2 to 18 years. One of the participants has also participated in the first validation iteration in May 2022. This participant's data are included in the data analysis as no large training effects are expected.

3.1.5. SCHEDULE FOR EXERCISE EXECUTION

As foreseen in E-OCVM, the whole validation process was performed in iterative loops to allow the adaptation or improvement of simulation environment, scenario, and tools to the desired objectives. For real-time simulation, two sessions were performed.

The simulation procedure is scheduled into different sections:

- **Briefing session**: This session intended to introduce the validation activities objectives, plan, organisation as well as the concept ideas and supporting tools being tested to the ATCOs participating to the trials.
- **Training session**: A training session was conducted for each ATCO at the very beginning of each trial before the simulation runs to familiarize the ATCOs with the simulation environment.
- **Simulation run 1 (60%):** a first simulation run of about 45 minutes was conducted with 60% share of 4D-FMS aircraft. After this run, a post-run questionnaire was answered by the ATCOs.
- **Simulation run 2 (30%)**: a second simulation run of about 45 minutes was conducted with 30% share of 4D-FMS aircraft. After this run, a post-run questionnaire and post-exercise questionnaire were answered by the ATCOs.
- **Explorative simulation run 3 (60%)**: A **debriefing** and an **explorative** simulation run were conducted to obtain more in-depth feedback about the system from the ATCOs. This took place after the final tailored questionnaire. During the explorative simulation run, individual components (ghosts, target windows and final distance indicator) were deactivated and activated one at a time. Participants

received some time to test the system when one of the components was deactivated and were asked how this affected their work as an opening question. This was followed by detail questions about each tool.

- **Simulation run 4 (80%):** a fourth simulation run of about 20 minutes was conducted with 80% share of 4D-FMS aircraft. No questionnaire is answered after this run. The aim of this short session was to collect ATCO feedback when the percentage of 4D-FMS aircraft is quite high. The idea behind was to push the limits of the feasibility of such concept. The evaluation of this run will be then made more based of the collected feedback during debriefing rather than from the simulation data log or questionnaire.

For the first iteration, the debriefing was conducted after all simulation runs, including the explorative one, were completed. For the second iteration, the debriefing took place after the first and second simulation run and was combined with the explorative simulation run. The debriefing questions were modified from the first to the second iteration. The feedback from the debriefings and the explorative simulation run was summarized in a qualitative manner.

3.1.5.1 SESSION 1 (MAY 2022)

The first session was performed from May 16th 2022 to May 25th 2022. Each day was organized as per agenda below.

Table 8. Validation activities agenda (session 1).

Time	Activity
08:15 – 08:30	Arriving
08:30 – 09:00	Briefing
09:05 – 09:35	**Training**
09:35 – 09:40	Post-training questionnaire
09:40 – 09:50	Short break
09:50 – 10:40	**Simulation run 1 (60%) + ISA**
10:40 – 10:45	Post-run questionnaire
10:45 – 10:55	Short break
10:55 – 11:45	**Simulation run 2 (30%) + ISA**
11:45 – 12:00	Post-run questionnaire, post-exercise questionnaire
12:00 – 13:00	Lunch
13:00 – 14:50	**Explorative simulation run 4 (60%)**
14:50 – 15:10	Debriefing

3.1.5.2 SESSION 2 (SEPTEMBER 2022)

The validation activities were performed from 5th to 9th of September 2022. Each day was organized as per agenda below:

Table 9. Validation activities agenda (session 2).

Time	Activity
08:15 – 08:30	Arriving
08:30 – 09:00	Briefing
09:05 – 09:35	**Training**
09:35 – 09:40	Post-training questionnaire
09:40 – 09:50	Short break
09:50 – 10:40	**Simulation run 1 (60%) + ISA**
10:40 – 10:45	Post-run questionnaire

10:45 – 10:55	Short break
10:55 – 11:45	**Simulation run 2 (30%) + ISA**
11:45 – 12:00	Post-run questionnaire
12:00 – 13:00	Lunch To be reviewed/ completed by
13:00 – 14:50	**Explorative simulation run 3 (60%) + Debriefing**
14:50 – 15:10	Post-run questionnaire, post-exercice questionnaire
15:10 – 15:20	Short break
15:20 – 15:45	**Simulation run 4 (80%) + ISA**
15:45 – 16:00	Debriefing

3.1.6. TRAFFIC SAMPLE

REAL-TIME SIMULATION

The airport selected for the airport-to-airport connection is Munich Franz-Josef-Strauß airport (EDDM). The traffic sample has been chosen to create a Mid Complexity/Mid Density environment. The scenario targeted 40 arrival movements and reduced departure movement (around 6 Movements). The traffic mix orientated on typically EDDM traffic situation in 2022. Since currently the share of aircraft with advanced FMS varies widely among airlines, the amount of 4D-FMS aircraft included in the scenario is used as decisive parameter to distinguish the scenarios. Table 10 provides an overview of the five developed simulation scenarios and their composition.

Table 10. Simulation scenario composition and overview.

Scenario ID	Total ARR	% of 4D-FMS ARR	Traffic Sample	% of ARR Heavy	Time
R[1]	18-22	0	2021	0	-
T	20	25	2019	13	14:00-14:45
S30	40	30	02.04.2022	23	07:00-08:00
S60	40	60	03.03.2022	19	09:00-10:00
S80	40	80	01.04.2022	3	18:00-19:00

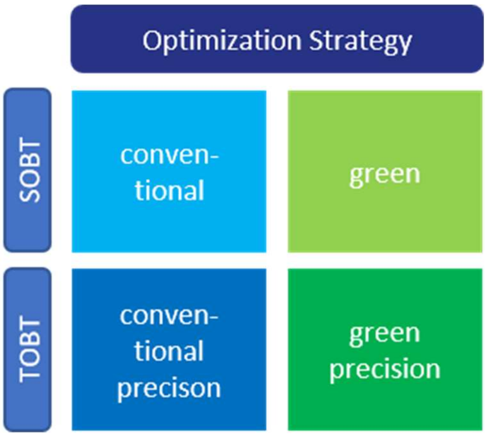
AUTOMATIC SIMULATION

Three different traffic scenarios with a length of one hour were used for the evaluation of the trajectories. A low traffic density scenario with 23 aircraft and a medium traffic density scenario with 36 aircraft were based on real traffic data (with two hours of traffic matched to one hour for the medium density scenario). A heavy traffic density scenario with 45 aircraft was designed artificially, but with a comparable traffic mix. All scenarios contained roughly an equal number of departures and arrivals. The flight plans for the departing aircraft contained either SOBTs or TOBTs, which could be used by the SMAN trajectory

[1] Data taken from OpenSky database [OpenSky 2023].

generator. For each scenario, two separate configurations of the flight plans were used. The default configuration contained SOBTs that were grouped in five-minute blocks, as is often the case with regular static flight plans. This means that for each five-minute block, it was possible that multiple departures were scheduled with identical SOBTs, leading to potential trajectory conflicts during pushback or the subsequent taxi phase.

In a second precision configuration, the flight plans were manually edited to simulate dynamically allocated, precise TOBTs, leading to a potentially lower risk of initial trajectory conflicts that need to be solved. The trajectory calculation algorithm has been adapted with a green optimization strategy, with a focus to reduce the number of holds during taxi. This is compared to a conventional taxi trajectory optimization strategy, that also generates conflict-free trajectories, but is more likely to use holds to solve conflicts. Figure 14 shows an overview of the four different combinations of planning times and optimization strategies.



	Optimization Strategy	
SOBT	conven-tional	green
TOBT	conven-tional precison	green precision

Figure 14. Overview of the optimization strategies used for the taxi trajectory calculation in combination with the different planning times that are used to calculate taxi trajectories.

3.2. VALIDATION EXERCISE EXE-002 – HC AND PILDO LABS

3.2.1. PLATFORM USED IN THE SIMULATION

The validations ran on MATIAS BEST simulator of HungaroControl. It has the same software as in the OPS room (i.e. MATIAS main ATM system), it supported the ecological validity of the validation.

1-1 sectors have been simulated within the Budapest Approach (1 Executive Controller (EC) and 1 Planner Controller, (PC), in two independent circuits. This way more participants could test the MergeStrip and could engage in the debriefing sessions at the end of the day. The two sectors were completely separated and did not use the same traffic sample at the same time. Pseudo-pilots played the role of the pilots, whilst feeders took care of the surrounding flights.

Figure 15. MATIAS-BEST simulator at HungaroControl. 2 Approach positions (EC, PC) have been measured at the same time.

The currently operational MergeStrip (reference) and the one that is developed by Pildo Labs (solution) has been integrated in the simulation environment, feeding the tool with data generated by MATIAS-BEST simulator. Each ATCO has been provided with its own MergeStrip client position.

3.2.2. OPERATIONAL ENVIRONMENT AND NEW TOOL

The exercise used the Budapest TMA, SIDs/STARs, TMA entry/exit and FRA intermediate points that have been re-designed and implemented in January 2020 (Figure 16).

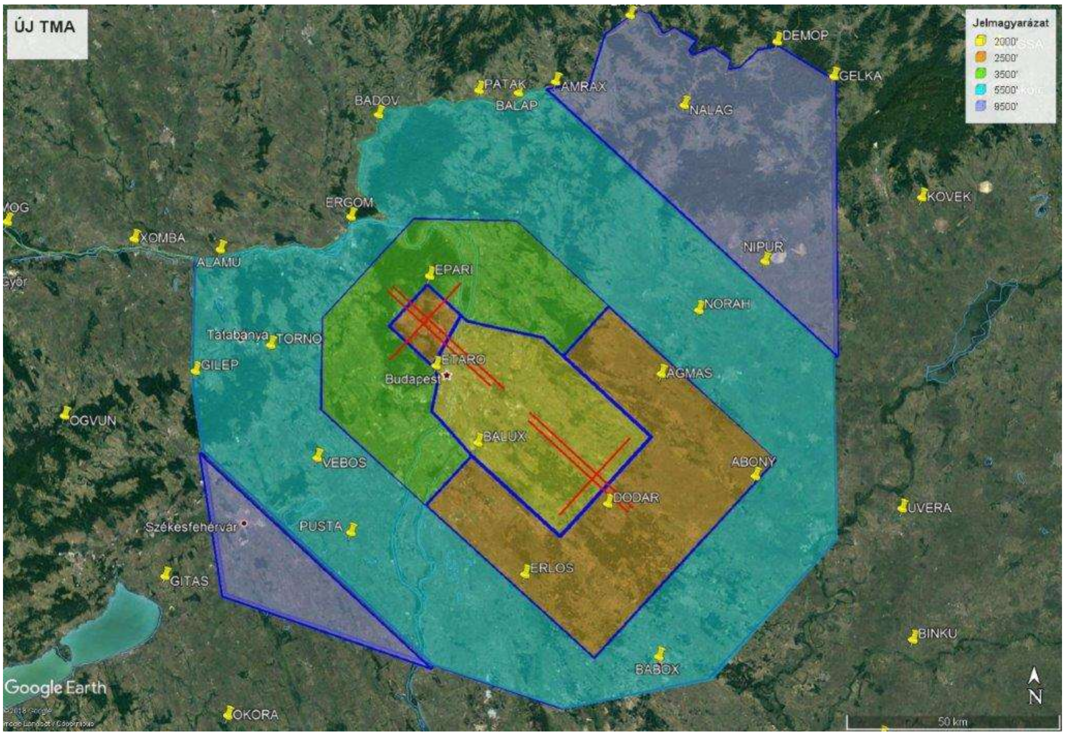


Figure 16. Chart of the new Budapest TMA (implemented on 30 January, 2020). The legend on the right side shows the height of each TMA sub-parts.

3.2.3. ROLES & RESPONSIBILITIES IN THE EXERCISE

The following roles are performed during the real-time simulation:

- **ATCOs**: Act as EC and PC during the Validation runs. Provided feedback about perceived workload, situational awareness, usability questions via questionnaires after the runs as well as during the debriefings.
- **Pseudo Pilots**: Provide radio communication to ATCOs during all runs and change aircraft trajectories on dedicated interfaces accordingly.

- **Validation Experts (Human Factors and Safety experts)**: Preparing and conducting data gathering, and analysing during and after the runs and contributing to the exercise report.
- **Technical Experts:** Preparing the platforms and scenario files, ensuring proper operation of the platform and all systems necessary for the simulation, supporting the gathering and analysis of quantitative data.
- **Exercise Lead:** Ensure timely and coordinated conduction of all runs, coordinate preparation and analysis of validations. Facilitate debriefing sessions.
- **Project manager:** Recruit participants, oversee the coordination between the project members, prepare the administrative materials (e.g. consent forms)

3.2.4. CONTROLLERS' BACKGROUND AND ROLES

All of the 6 ATCOs who participated in the sessions are active, licensed air traffic controllers in the Budapest Approach Unit. During the trials, they were responsible for standard arrivals and some departures. Only the Exercise Leads (2 APP ATCOs) were familiar with the new MergeStrip, as they were core team members for its design. The 6 ATCOs who participated in the simulations know the MergeStrip used in the reference runs very well.

The idea behind recruiting the participants was the following:

- 4 ATCOs could simulate at the same time, in two independent circuits (2 sectors with 2-2 ATCOs)
- The validation team partly changed between the first and the second iteration. 2 ATCOs remained the same to enable continuity (i.e. they remembered how the system behaved in the first iteration and could evaluate whether it has been implemented and improved), and the other two ATCOs were newcomers who could judge the system without having a preconception about the past.
- This equals in 6 APP ATCOs in total who participated in the validation session

3.2.5. SCHEDULE FOR EXERCISE EXECUTION

The autumn validation session has focused on the what-if functionality. The session has been separated into two iterations to make sure that the feedback obtained in the first iteration can be integrated into the MergeStrip and ATCOs can test the software again, with different traffic samples.

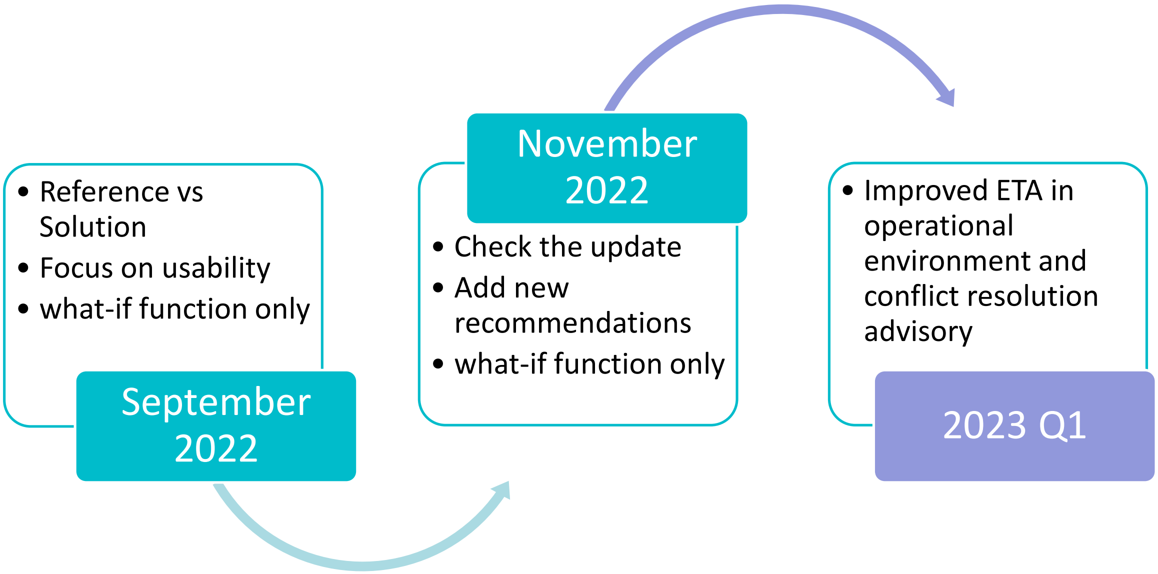


Figure 17. Aim of the iterations.

The aim of first iteration was to already analyse the difference between the reference and solution scenarios (i.e. original and new MergeStrip software). For that the team used four traffic samples. The two main variables were runway direction (13 vs 31) and runway direction change (13→31 or 31→13).

Table 11. Overview of the used runway directions and the associated scenario IDs.

Runway direction	Scenario ID
13	105-OB
31	106-OB
31->13	203-RF
13->31	204-RF

The simulation procedure is scheduled into different sections:

- **Briefing session**: this session intended to introduce the validation activities objectives, plan, organisation as well as the concept ideas and supporting tools to be tested to the ATCOs participating to the trials.
- **Training session**: A training session was conducted for each ATCO at the very beginning of each trial before the simulation runs to familiarize the ATCOs with the new MergeStrip functionalities.
- **Simulation runs** and Post-run questionnaire with workload and situational awareness questions.

Table 12. Scenarios in the first iteration (Day 1 and Day 2).

Day 1				CIRCUIT 1 (pos01 and 02)	CIRCUIT 2 (pos03 and 04)
Start	Duration	End	Activity		
9:00	0:15	9:15	Briefing		
9:15	0:30	9:45	Training (13->31)		
9:45	0:45	10:30	Run 1	NEW MS: 105-OB	REF MS: 106-OB
10:30	0:20	10:50	Questionnaire, break		
11:45	0:45	12:30	Run 2	NEW MS: 106-OB	REF MS: 105-OB
12:30	0:10	12:40	Questionnaire		
12:40	1:00	13:40	Lunch		
13:15	0:45	14:00	Run 3	NEW MS: 203-RF	REF MS: 204-RF
14:00	0:10	14:10	Questionnaire, break		
14:10	0:45	14:55	Run 4	NEW MS: 204-RF	REF MS: 203-RF
14:55	0:30	15:25	Questionnaire, debrief		

Day 2				CIRCUIT 1 (pos01 and 02)	CIRCUIT 2 (pos03 and 04)
Start	Duration	End	Activity		
9:00	0:45	9:45	Run 1	NEW MS: 106-OB	REF MS: 105-OB
9:45	0:20	10:05	Questionnaire, break		
10:05	0:45	10:50	Run 2	NEW MS: 105-OB	REF MS: 106-OB
10:50	0:20	11:10	Questionnaire, break		
11:10	0:45	11:55	Run 3	NEW MS: 204-RF	REF MS: 203-RF
11:55	1:00	12:55	Lunch		
12:55	0:45	13:40	Run 4	NEW MS: 203-RF	REF MS: 204-RF
13:40	0:45	14:25	Questionnaire, debrief		

The aim of the second iteration was to fully focus on the new MergeStrip and

- to check whether the feedback given has been properly integrated into the software,
- to collect new recommendations for future improvements.

Table 13. Scenarios in the second iteration (Day 1 and Day 2).

Day 1				CIRCUIT 1 (pos01 and 02)	CIRCUIT 2 (pos03 and 04)
Start	Duration	End	Activity		
7:30	0:15	7:45	Briefing		
7:45	0:30	8:15	Training	MS-MULTI-203-PZ (RWY change) 13->31	
8:15	0:45	9:00	Run 1	201-RF runway direction: 13	202-RF runway direction: 31
9:00	0:20	9:20	Questionnaire, break		
9:20	0:45	10:05	Run 2	202-RF runway direction: 31	201-RF runway direction: 13

10:05	0:10	10:15	Questionnaire, break		
10:15	0:45	11:00	Run 3	203-PZ (runway change) 13->31	204-PZ (runway change) 31-> 13
11:00	0:10	11:10	Questionnaire, break		
11:10	0:45	11:55	Lunch		
11:55	0:45	12:40	Run 4	204-PZ (runway change) 31-> 13	203-PZ (runway change) 13->31
12:40	0:30	13:10	Questionnaire, debrief		

Day 2				CIRCUIT 1 (pos01 and 02)	CIRCUIT 2 (pos04 and 05)
Start	Duration	End	Activity		
7:30	0:15	7:45	Briefing		
7:45	0:45	8:30	Run 1	305-PZ (runway change) 13->31	306-PZ (runway change) 31-> 13
8:30	0:20	8:50	Questionnaire, break		
8:50	0:45	9:35	Run 2	306-PZ (runway change) 31-> 13	305-PZ (runway change) 13->31
9:35	0:20	9:55	Questionnaire, break		
9:55	0:45	10:40	Run 3	301-OB runway direction: 13 heavy	302- OB runway direction: 31 heavy
10:40	0:10	10:50	Questionnaire, break		
10:50	1:00	11:50	Lunch		
11:50	0:45	12:35	Run 4	302- OB runway direction: 31 heavy	301-OB runway direction: 13 heavy
12:35	0:45	13:20	Questionnaire, debrief		

The new validation session, scheduled to be held early next year will use the updated version to make the final assessment by adding the improved ETA.

3.3. DEVIATIONS FROM THE PLANNED ACTIVITIES

3.3.1. DEVIATIONS WITH RESPECT TO THE VALIDATION PLAN

- **EXE-001 – DLR**

The traffic load (medium vs. high) was not varied. Only the medium traffic load is tested. The full capacity scenario was not conducted.

The traffic distribution was varied at different levels than envisaged in the VALP: 30% 4D-FMS vs. 60% 4D-FMS vs. 80% 4D-FMS instead of 0% 4D-FMS vs. 40% 4D-FMS vs. 75% 4D-FMS

Due to technical issues, the 80% 4D-FMS equipped aircraft scenario was not conducted for all participants. In the first iteration, the 80% simulation run was not conducted at all and instead replaced with a training scenario (for X participants). These data were excluded from data analysis. During the second iteration, the 80% simulation run was conducted for all participants after the other simulation runs and the debriefing have been conducted. It was not run in its full length, but for around 20 minutes only. ISA ratings were obtained during the 80% scenario and participants did not fill out any post-run questionnaires afterwards.

It was initially planned to involve 10 ATCOs in the trials but unfortunately only 5 ATCOs were involved because of a shortage of available controllers to participate in this validation exercise.

EXE-002 – HC & PILDO LABS

Real-time Simulation

- The traffic load (medium vs. high) was only manipulated in the second iteration. Scenarios with runway change were considered medium density traffic, whereas the simple runway direction scenarios were more difficult.
- Only the what-if functionality has been tested in the first session. The remaining functionalities described in the Validation Plan will be simulated in the next year's session (i.e. improved ETA and conflict resolution advisory). The reason for this was that after the first iteration the team decided to put more focus on improving the what-if functionality, based on the feedback the ATCO gave. The second iteration enables the team to check what has been developed and how that compared to what they were looking for in the software. As the focus of the validations were the software capabilities and the What-if function, the execution of a complete, industry standard safety analysis was not possible. The safety assessment will be carried out in the second validation iteration in 2023.

Shadow Mode

- Originally no passive shadow validations were planned, but only Real-Time simulations in simulator environment (which is identical to main system, MATIAS), and these RTS sessions took place in September and November 2022). During the validation activity it was realised that although technically feasible, the ML based functionalities were not worth to be tested in a simulation environment. To achieve more realistic outcomes, it was decided to move the validation exercise to the OPS room and validate the remaining functionalities instead of sticking to RTS with the 'What-if' functionality with the original scenarios (e.g. abnormal scenario and degraded mode). At the same it meant, that Pildo and HungaroControl went beyond the testing needs prescribed by Call for Proposals and undertook in the Grant Agreement (TRL-4 level).
- Hard as the developers tried, in the initial phase of the validation, MergeStrip could not become stable enough to confidently judge its usability, as it often froze during the validation sessions. As a means of mitigation, the MergeStrip service was restarted every 30 minutes, and the service became stable, and was fully functional. However, this had a huge impact on the answers received from users in the questionnaires and the final workshop whose results are presented in this document.
- Cost-effectiveness could not be measured in a passive shadow mode validation and thus not part of this Validation Report. Similarly, workload and situational awareness related objectives are not within the scope of the passive shadow mode.

4. VALIDATION RESULTS

4.1. VALIDATION EXERCISE EXE-001 – DLR

The results are sorted by the KPAs environment, operational efficiency and safety.

4.1.1. ENVIRONMENT

In this KPA, it is assessed whether the operations supported by new airspace design and supporting tools will have a positive impact on the environment.

Table 14. EXE-001 - Environment KPA results.

Criteria ID	Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
ENV–GREAT–CRT-01-10	Less fuel burned as average of complete traffic scenario compared to reference scenario	*Results by project partner UPM indicate improvements in fuel burn with increased percentage of 4D-FMS aircraft. As the reference traffic data from OpenSky did not include thrust values, no baseline comparison could be completed. Results from the automatic simulation of taxi operations indicate a significant reduction of fuel burn through a reduction of holds during taxi phase.*	OK

4.1.1.1 AVERAGE FUEL BURN PER FLIGHT IN TMA

The average fuel burn per flight has been computed by project partner UPM, based on an interpolation of fuel burn of aircraft engines in different thrust settings, as validated in the ICAO engine emission database, in combination with calculated thrust settings by the NARSIM simulator used in the HITL trials at DLR.

4.1.1.2 AVERAGE FUEL BURN PER FLIGHT DURING TAXI

For the evaluation of fuel burn during the taxi phase, the automatic simulation runs using the DLRs SMAN were evaluated. The main focus was on eliminating conflicts during the taxi phase, which lead to holding times and consequently to increased fuel burn because of idling time and accelerating with higher thrust settings. The evaluation is based on five automated simulation run for each combination of traffic scenario and optimization strategies/planning times configuration (Figure 14).

Figure 18 shows the analysis of the average number of stops during taxi within the automated simulation runs. The horizontal axis presents three distinct sections for the low, medium and high traffic density scenarios. For each traffic scenario, four different configurations have been simulated. As can be seen, the low-density traffic scenario did not produce a large number of holds in either configuration. In medium and high-density scenarios, the number of holds rises and shows a clear reduction in number of holds when using the green optimization strategies for both SOBT and TOBT-based planning.

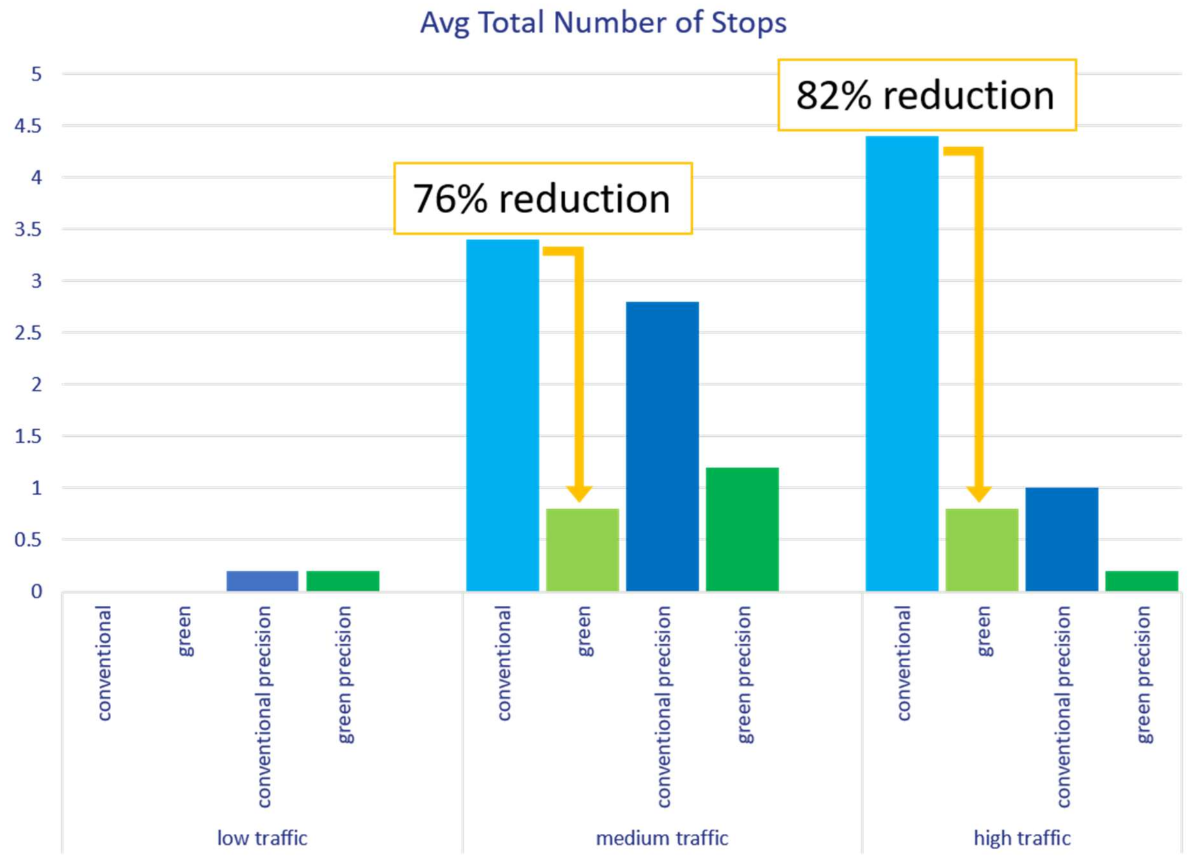


Figure 18. Total number of stops averaged over 5 simulation runs per configuration.

It is hard to exactly quantify the impact of the reduction of holds on fuel, because performance and fuel consumption data in both BADA and ICAO models is not modelled precise enough to take the additional fuel burn because of holds into account. Based on previous research conducted on flight data recordings [Grier 2015], it can be assumed that holds account for up to 18% of fuel burn during taxi. Based on this, it can be assumed that up to 14% of fuel burned during taxi can be saved by using optimized green taxi trajectory algorithms in SMAN.

4.1.2. OPERATIONAL EFFICIENCY

Within this KPA, it is intended to assess the efficiency of the new system (airspace structure and controller assistance tools). The aim would be to check if the flown distance within an observation horizon as an average of complete traffic scenario is reduced compared to baseline. The results are detailed in Chapter 4.1.1.1 and resumed in Table 15.

Table 15. EXE-001 - Operational Efficiency KPA results.

Criteria ID	Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
OPE–GREAT–CRT-02-10	The distance flown is reduced compared to reference scenario.	The results show a shortage of the flown distance. Even through the relatively small difference, it could be seen that the results were always lower as reference values.	OK

4.1.2.1 FLIGHT DISTANCE IN TMA

An analysis of the traffic distribution is carried out in order to determine appropriate research horizon and measure the length of the travelled trajectories, which has also contributed to the estimation of fuel consumption reduction, prepared on the basis of OpenSky [OpenSky 2023] and BADA data [Nuic 2010].

In Figure 19, the vertical axis presents for arriving aircraft the average distance flown within the radius of 100 NM to the Aerodrome Reference Point (ARP). The horizontal axis presents the results of validation trials executed by five ATCs (C1 - C5) testing two traffic scenarios, differing with distribution of 3D-FMS and 4D-FMS flights, where 30 and 60 corresponds respectively to 30% and 60% of the FMS air traffic operations. The results obtained in conducted validation activities have been marked as orange line. They can be directly compared with real traffic data marked in blue, where the distance flown to the ARP has been calculated as an average based on 10 hours of arrival traffic in Munich with the same amount of traffic flow extracted from the OpenSky database.

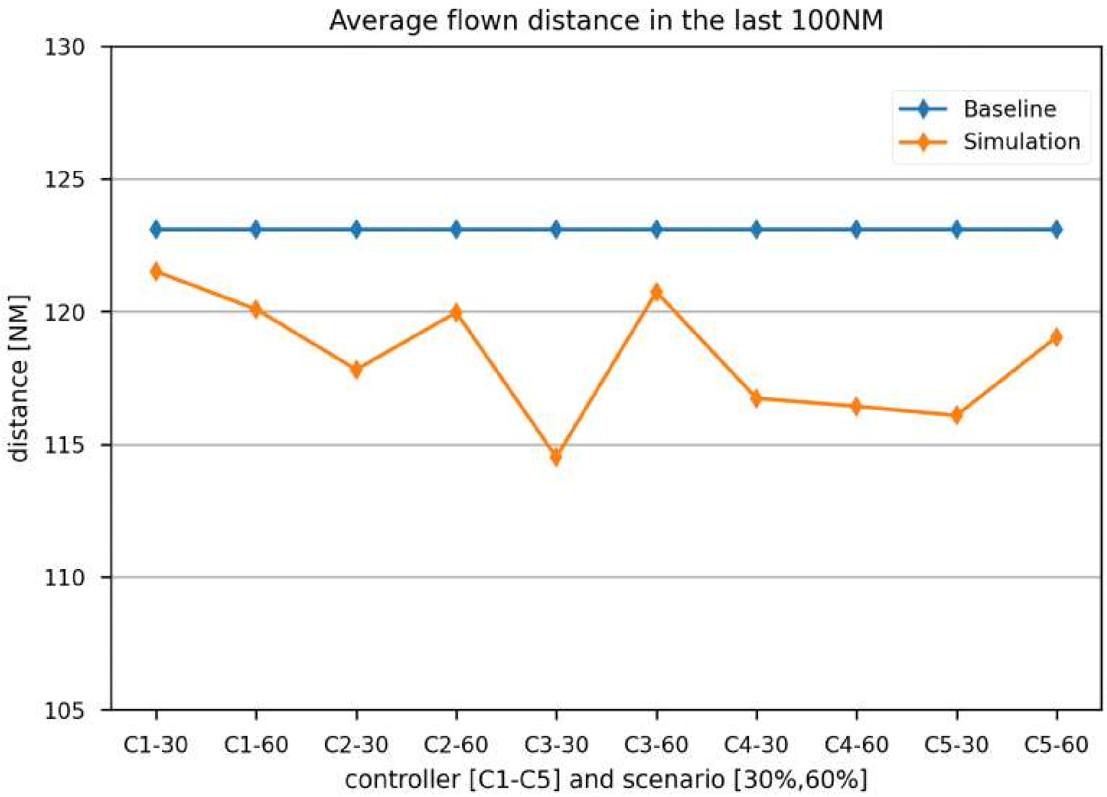


Figure 19. Flight trajectories results obtained under the simulation conditions and compared with real traffic reference values.

Treating that as reference, it can be observed that the introduction of innovative airspace structure, new FMS procedures and ATCs supporting systems results in the shortage of flight distance across all ATCs and all scenarios. Although this is statistically not significant, varying between 114 a 121 NM, in each case the results were lower as reference values, what also indicates that application of this type of solution has proved to be effective.

Figure 20 presents the aircraft’ flown distances as cumulative occurrence curve divided into 5-NM lengths, where again the orange line refers to the trial values and the blue line presents the baseline data. The Analysis points out that, in the simulated solution, the number of flights covering shorter distances is slightly higher in comparison to conventional flight operations carried out today. That is particularly evidenced by the first two peaks observed in Figure 20, which are significantly higher than the baseline and thus indicate that well over half of scheduled flights (sum of 68,3%) arrived at the airport in the exceptionally small range of 100-115 NM. In the same range for the baseline data that value was equal to 36,9%, almost twice as low compare to the solution scenarios. In addition to that, the real traffic data show that a significant amount of flights (corresponding to the 25% of occurrence), needed a distance of 125 NM to reach the airport. This noticeably reinforces the effect of the environmental benefits of the proposed concept as well.

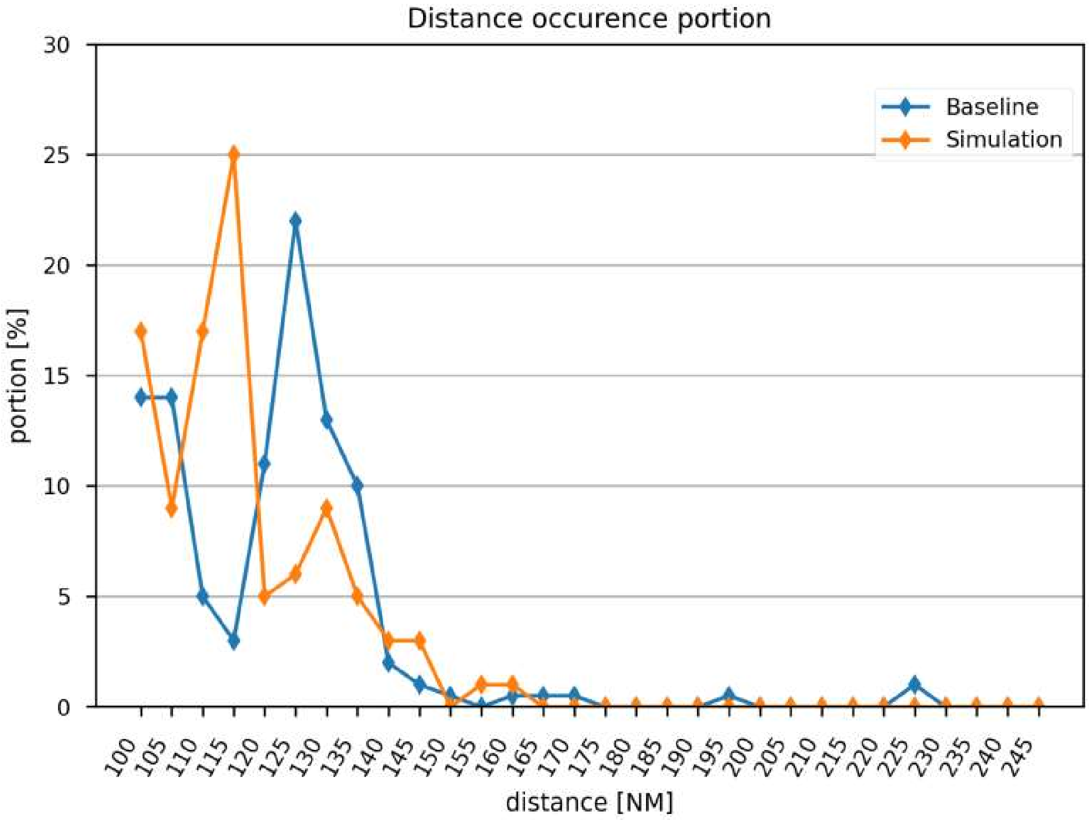


Figure 20. Cumulative occurrence curve for flight distance in relation to baseline data and result obtained in validation trials.

The results of the validation of the new airspace structure for the separation of approach flows into conventional traffic and optimized profile descents show a reduction of the average approach distance by six Nautical miles depending on the proportion of direct approaches and thus the potential of the developed solutions to reduce greenhouse gas emissions from air traffic. However, the quantitative assessment of these emissions is difficult because the traffic simulator can calculate continuous descent approaches in principle, but cannot optimise them individually for individual aircraft types, as envisaged in the project concept. However, the 4D trajectories generated provide the basis for an environmental assessment of the approaches. The consequence of this rudimentary trajectory calculation is that the simulations of greenhouse gas emissions determined for the future scenarios do not provide unambiguous results. As a result, not all project elements allow a clear statement on their climate impact.

4.1.3. SAFETY

4.1.3.1 SAFETY PERFORMANCE

Within this KPA, the aim was to check if the new procedures and system functions proposed by the solution are safe in normal situations. it was intended to assess this KPA in two ways:

- Objectively: Through the number of separation infringements extracted from the simulation log and;
- Subjectively: From the ATCO perspective through questionnaires and debriefing.

The success criteria SAF–GREAT–CRT-09-10 was then spitted in 2 sub-criteria a and b. The results are collected in Table 16. It should be noted that EXE-001 covered only the normal situation. Therefore, the criteria SAF–GREAT–CRT-09-20 is not addressed. Given that the proposed tools are early prototypes, the main focus here was to check if they are providing the expected support to controllers.

Table 16. EXE-001- Safety performance results.

Criteria ID	Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
SAF–GREAT–CRT-09-10a	Procedures and system functions are safe in normal situations.	Mean agreements to the tailored questions regarding safety and confidence were neutral to positive. Nevertheless, ATCOs also voiced some potential safety risks.	OK
SAF–GREAT–CRT-09-10b	There are no critical separation infringements	To be checked in the simulation log	N/A
SAF–GREAT–CRT-09-20	Procedures and system functions are safe in abnormal situations.	The abnormal situation is not covered by the exercise	N/A

Controllers answered questions related to the perceived level of safety in both validation iterations in the final tailored questionnaire post-exercise. The results are summarized in Table 16. The questions asked can be seen in Figure 21 and Figure 22. The first three questions assessed the safety level perceived from the ATCO perspective. The last three questions aimed to check if the proposed tools helped to maintain/ improve safety level. The controllers rated the statements from 1 (strongly disagree) to 5 (strongly agree). Mean ratings were calculated. Any mean rating less than 3 for the first three (first iteration)/the first four (second iteration) questions would be an indication of unsafe operation.

FIRST ITERATION

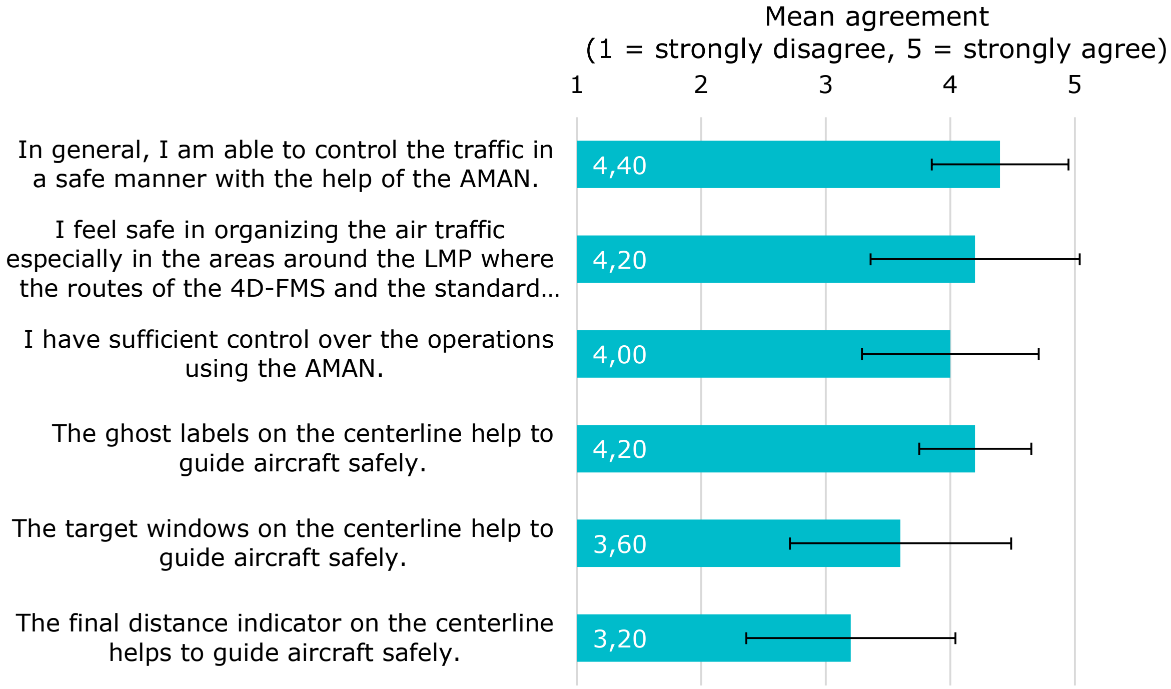


Figure 21. Mean agreement to tailored statements regarding the perceived safety. Error bars represent standard deviations.

As can be seen from Figure 21, all statements regarding the perceived safety received a mean rating always higher than 3 (neither agree nor disagree). From the first three items, it becomes apparent that the ATCOs seemed to feel in control and safe in controlling the traffic. The statements regarding the target windows and the final distance indicator received the lowest mean agreements. This indicates limited benefits of the target windows and the final distance indicator for safety. In contrast to that, the ghosts seemed to be more helpful regarding safety.

During the debriefing, one participant named the departures climbing too slowly as a safety critical situation since the departure traffic is handled by the DMAN and then the ATCO could not interact with it in case of conflict with arrivals. The remaining four participants reported no safety critical situations except for technical issues related to the simulation. All five participants affirmed that they felt safe in organizing the traffic around the LMP.

However, some **critical comments** regarding safety were raised:

- The TMA was perceived as too big. The ATCO needed to zoom in and zoom out several times, which could be safety critical when events happen outside of the displayed area.
- The AMAN system computes and proposes the optimized arrivals sequence. Unless s/he otherwise decides, the ATCO does not have to think about the sequence anymore. This could potentially become a safety issue when the controller needs to take over for some reasons.

SECOND ITERATION

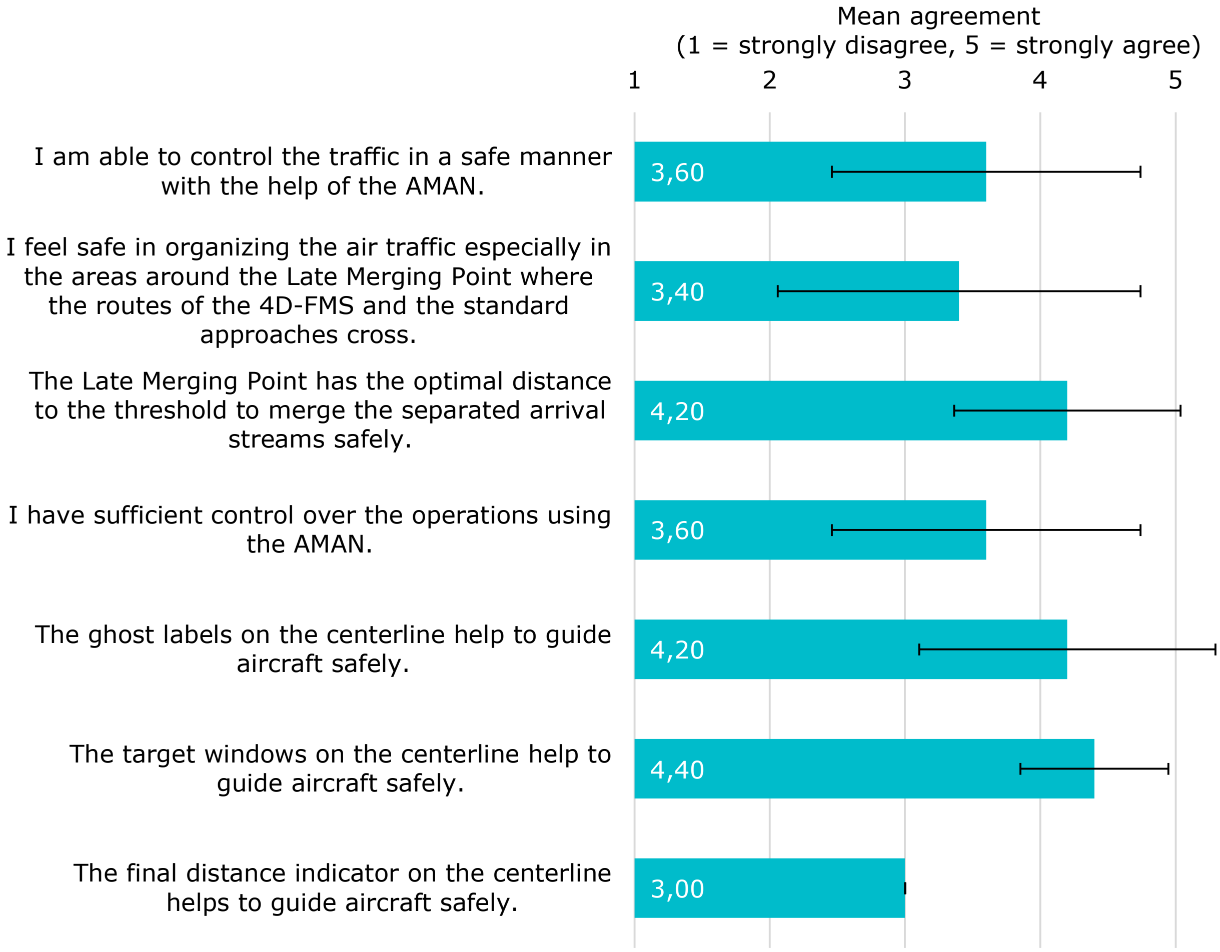


Figure 22. Mean agreement to tailored statements regarding the perceived safety. Error bars represent standard deviations.

Figure 22 shows that the tailored statements regarding perceived safety were given an average rating of 3 (neither agree nor disagree) at the minimum. From the first four items, it can be seen that the ATCOs reported safe operations within the new airspace design overall but there seems to be room for improvement. Even though the distance of the LMP to the threshold was overall rated as safe, the area around the LMP was identified as one area of improvement regarding safety. The standard deviation of the statement regarding the traffic around the LMP was rather high, indicating a wide distribution of the answers given. The last statement regarding the final distance indicator received an unambiguous rating of 3 (neither agree nor disagree), indicating that the final distance indicator was of limited helpfulness concerning safety. Compared to this, the ghosts and the target windows seem to have been more helpful.

Safety critical situations reported by the participants during the debriefing were related to the simulation, technical issues or lack of experience with the system, e.g.:

- Aircraft going opposite at the LMP. This happened because arrival sequences computed for both runways are independently calculated by the AMAN since the two runways were independent. One ATCO suggested to shift one of the LMPs for safer operations on both runways.
- Difficulty to judge distances (lack of measurement tools, unknown airspace…)

SUM UP

The Safety Performance questionnaires were slightly extended from the first validation to the second one, so that a summary statistical evaluation is not possible. However, it is noticeable that both Safety Performance indicators in terms of aircraft guidance around the LMP and the use of Ghosts were rated identically and the Final Distance Indicator was rated almost identically in both iterations.

At the same time, however, the assessment regarding the use of the AMAN in the 2nd iteration was significantly lower than in the first one. This was partly due to the fact that in one scenario two aircraft were simultaneously guided towards the adjacent LMPs and thus flew directly towards each other for a brief moment. On the other hand, some controllers missed a function for continuous distance measurement of two aircraft. In operational use, this tool is used to monitor the prescribed separation. This function was available during the trials, but it was only noticed by the controllers at a very late stage due to requests.

4.1.3.2 CONFIDENCE IN USING THE NEW TOOLS

FIRST ITERATION

No questions regarding confidence were included in the first iteration.

SECOND ITERATION

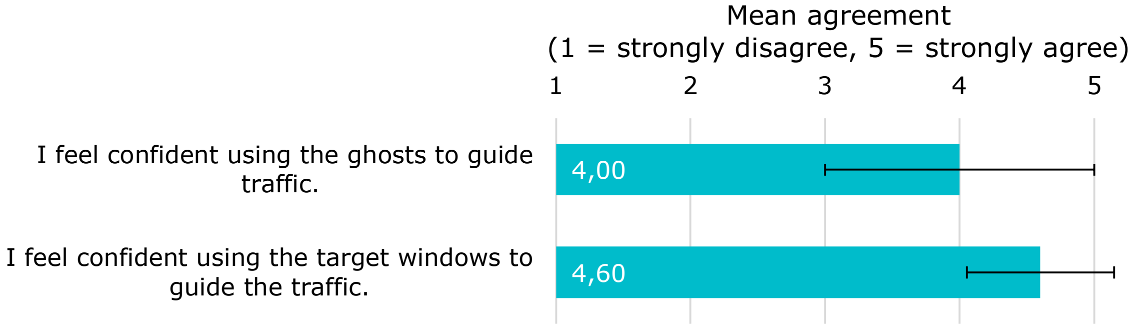


Figure 23. Mean agreement to tailored statements regarding Confidence. Error bars represent standard deviations.

Overall, participants agreed to feeling confident when using the ghosts and the target windows, see Figure 23.

During the debriefing, four participants stated that they felt generally confident. One participant reported that confidence was low at first but increased “exponentially” in the course of the day. This is understandable because normally it is necessary to sufficiently try a new tool to be able to trust it and use it.

4.1.4. HUMAN PERFORMANCE

4.1.4.1 WORKLOAD

Mental workload was assessed during the first and second iteration. Results are summarized in Table 17. ISA ratings were obtained every 5 minutes during the simulation runs through a touch screen display in order to assess mental workload from 1 (under-utilised) to 5 (excessive). Furthermore, the NASA-TLX without the subscale “physical demand” was administered post-run. Raw TLX scores ranging from 0 to 100 were calculated, with 0 indicating low demand and 100 indicating high demand. A mid-level mental workload is desirable, while more extreme values point to over- or underload.

Table 17. EXE-001 - Validation results for HUM– GREAT–CRT-05-10.

Criteria ID	Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
HUM–GREAT–CRT-05-10	The level of workload is within acceptable limits.	ISA ratings and raw NASA-TLX scores indicated no mental overload. However, ISA ratings in the 60% run of the first iteration and the 80% run in the second iteration pointed towards mental underload. Global raw TLX scores indicated low levels of mental workload compared with the observed global NASA-TLX scores according to Grier (2015) [Grier 2015] as well. One participant mentioned low levels of mental workload as a possible safety risk. Mental Workload could be expected to be lower in simulations than in real operations due to a more abstract representation of operations. Therefore, the observed mental workload was interpreted as acceptable, but it is strongly recommended to test mental workload in real operations.	OK

FIRST ITERATION

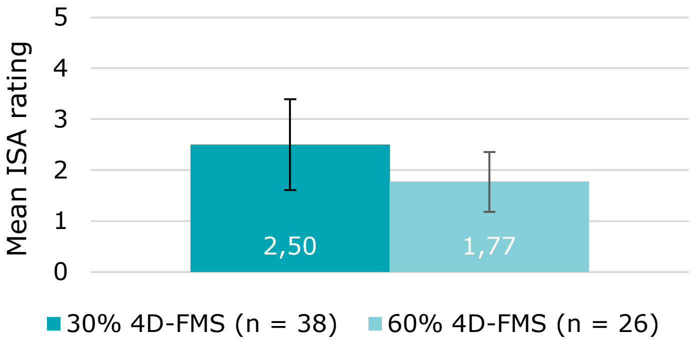


Figure 24. Mean ISA ratings in dependence of traffic distribution (30% vs. 60%) summarized over all participants and assessment times. Error bards represent standard deviations.

Figure 24 shows the mean ISA ratings for the 30% and the 60% run, calculated over all participants and assessment times. Descriptively, mean ISA ratings were higher in the 30% run than in the 60% run. For the 30% run, the mean ISA rating ranged between 2 (relaxed) and 3 (comfortable), indicating a slightly lower than mid-level mental workload. For the 60% run, the mean ISA rating was lower than 2 (relaxed).

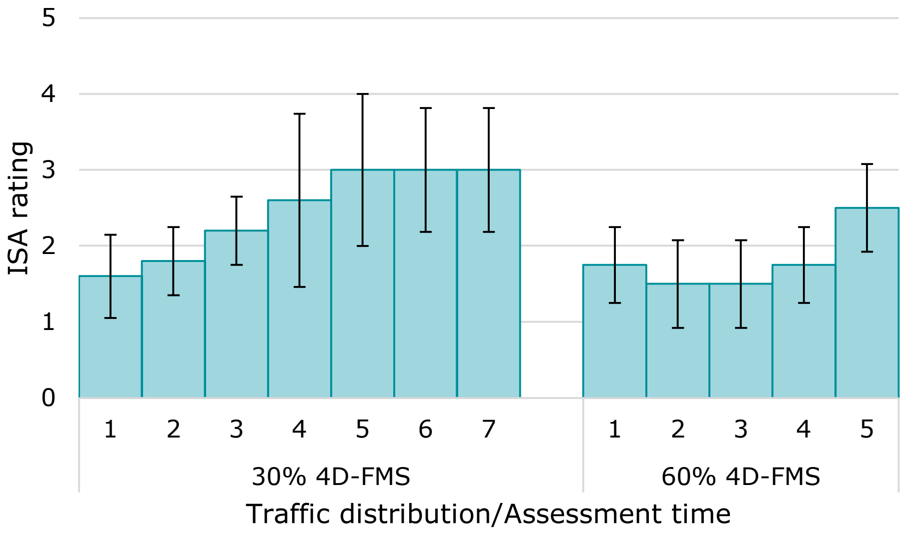


Figure 25. Mean ISA ratings in dependence of traffic distribution (30% vs. 60%) and assessment time (1 – 8). Error bars represent standard deviations.

Figure 25 shows the mean ISA ratings in dependence of traffic distribution and assessment time. Assessment times with less than four data points were not included. The mean ISA ratings ranged between 1 (under-utilised) and 3 (comfortable). For the 30% run, it can be seen that ISA ratings increased over the course of the run from M=1.60 (SD=0.55) to M=3.00 (SD=0.82). For the 60% run, the ISA rating remained more or less stable.

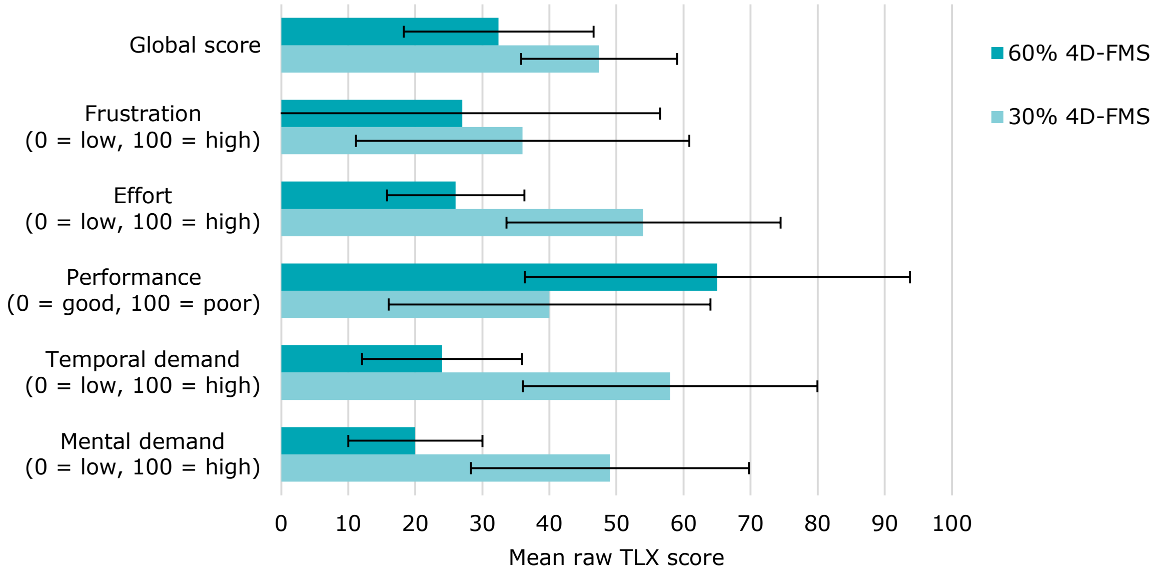


Figure 26. Mean Raw TLX scores in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

Figure 26 shows the mean raw TLX scores in dependence of the traffic distribution. For the global score and all TLX-subscales except for performance, raw TLX scores were higher in the 30% run than in the 60% run, indicating higher workload during the 30% than the 60% run on a descriptive level. This is in line with the ISA ratings. The mean score of the subscale performance was higher for the 60% run than the 30% run, i.e. participants rated their performance better in the 30% run than in the 60% run. This could be explained by the fact that in the 60% scenario, most of the traffic is handled by the system rather by the ATCOs. Standard deviations were especially high for performance and frustration, meaning there was a wide distribution of answers. According to Grier [Grier 2015], the mean global score for the 60% run (M=32.40, SD=14.15) fell below the 25th percentile of observed global NASA-TLX scores in the field of Air Traffic Control (ATC), while the mean global score for the 30% run (*M*=47.40, *SD*=11.63) fell below the 50th percentile of observed global NASA-TLX scores in the field of ATC. Here, it has to be considered that the subscale "physical demand" was excluded from the NASA-TLX in the validation exercises at hand. This limits the comparability with Grier's observed scores.

During the debriefing, two ATCOs reported that their workload was never outside acceptable levels. For two ATCOs, their experienced workload was too high in the beginning. This was attributed to inexperience with the system (working with a new system, new airspace, high traffic and vectoring…).

One participant stated that the system reduces workload because the ATCO does not have to think about the sequence anymore. However, this was also seen as a potential safety risk, cf. Section 0.

SECOND ITERATION

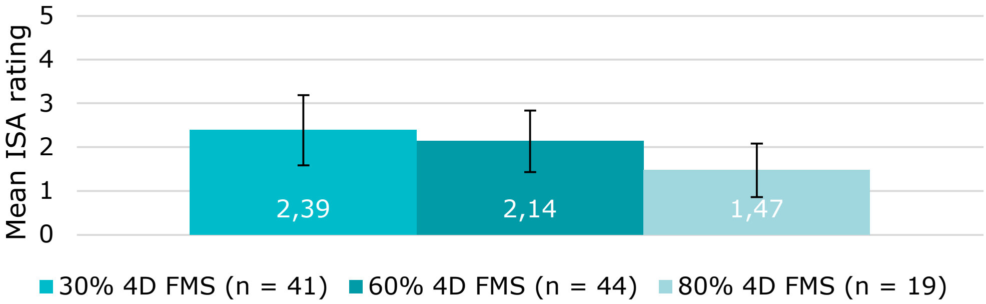


Figure 27. Mean ISA ratings in dependence of traffic distribution (30% vs. 60% vs. 80%) summarized over all participants and assessment times. Error bards represent standard deviations.

Figure 27 shows the mean ISA ratings for the 30%, the 60% and 80% run, calculated over all participants and assessment times of the two validation weeks. Descriptively, mean ISA ratings were highest in the 30% run, followed by the 60% run and then the 80% run. For the 30% and the 60% run, ISA ratings fell between 2 (relaxed) and 3 (comfortable), indicating a slightly lower than mid-level mental workload. For the 80% run, mean ISA ratings were below 2 (relaxed), pointing toward mental underload.

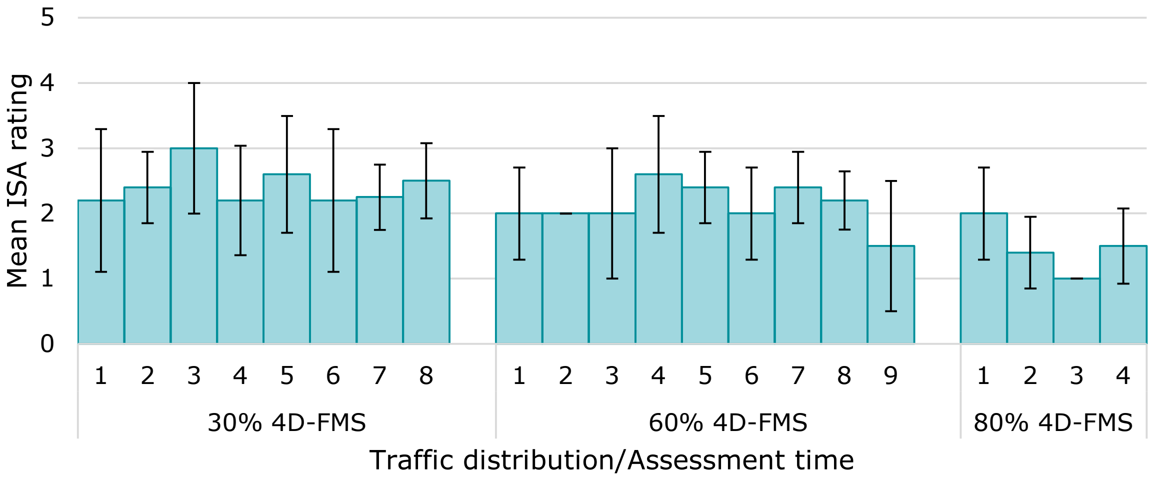


Figure 28. Mean ISA ratings in dependence of traffic distribution (30% vs. 60% vs. 80%) and assessment time (1 – 9). Error bars represent standard deviations.

Figure 28 shows mean ISA ratings in dependence of traffic distribution and assessment time. Assessment times with less than four data points were not included. Most mean ISA ratings ranged between 2 (relaxed) and 3 (comfortable). For the 80% run, the mean ISA ratings fell between 1 (under-utilised) and 2 (relaxed). It should be noted that the controllers' perceived metal workload was quite stable during each run with a slight increase around the middle of the scenario where normally the ATCO is handling more traffic compared to beginning/ end of the scenario.

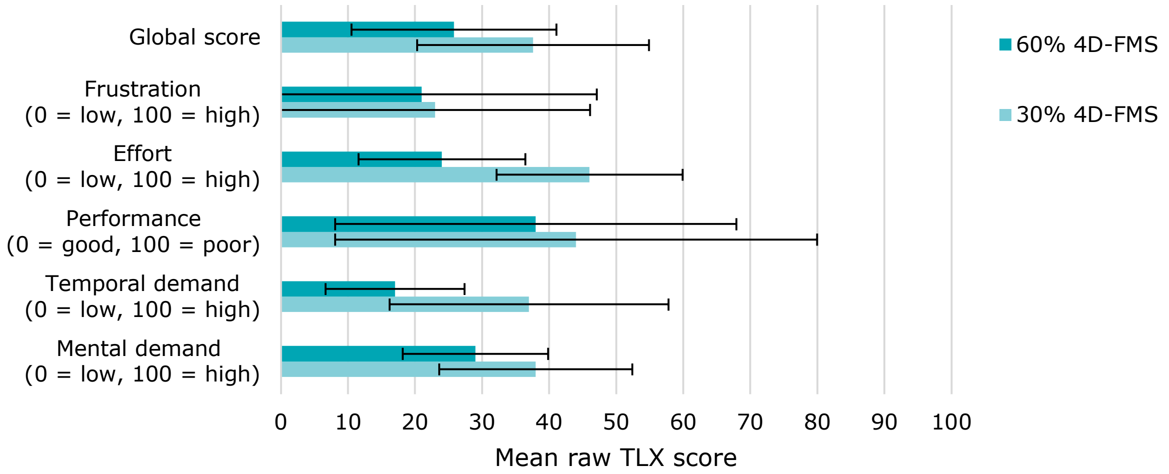


Figure 29. Mean Raw TLX scores in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

Figure 29 shows the mean raw TLX scores in dependence of the traffic distribution. The mean global score and all mean sub-scores were higher in the 30% run than in the 60% run, indicating a higher mental workload in the 30% run than in the 60% run on a descriptive level. This is in line with the ISA ratings and the first iteration, with the exception of the subscale performance. Again, standard deviations were especially high for the subscales frustration and performance. According to Grier, both mean global scores (*M*=25.80 [*SD*=14.27] for the 60% run and *M*=37.60 [*SD*=17.29] for the 30% run) fell below the 25th percentile of observed global NASA-TLX scores in the field of ATC [Grier 2015]. Here, it has to be considered that the subscale "physical demand" was excluded from the NASA-TLX in the validation exercises at hand. This limits the comparability with Grier's observed scores.

During the debriefing, two of five ATCOs reported that workload was never outside acceptable levels, two ATCOs reported low but acceptable workload in the 60% run. One of five ATCOs reported experiencing too low workload in the 60% run.

Another critical comment from one of five ATCOs during the debriefing was that the routes were too crowded and could result in an increase of workload.

SUM UP

The overall results of ISA and FISA show quite clearly that workload and the fatigue that often accompanies it decrease with increasing automation in approach guidance. Where workloads are still reported as above the mean in the 30% scenarios, they drop to a mean value of about two in the 60% scenarios and then go below this value in the 80% scenarios (Figure 30). Overall, however, the self-assessment of workload is quite high, and the dispersion between the individual test participants can also be seen clearly in all cases. This is due to the fact that the controllers had to take on three to four work positions alone in the trials, corresponding to a situation with a high traffic load.

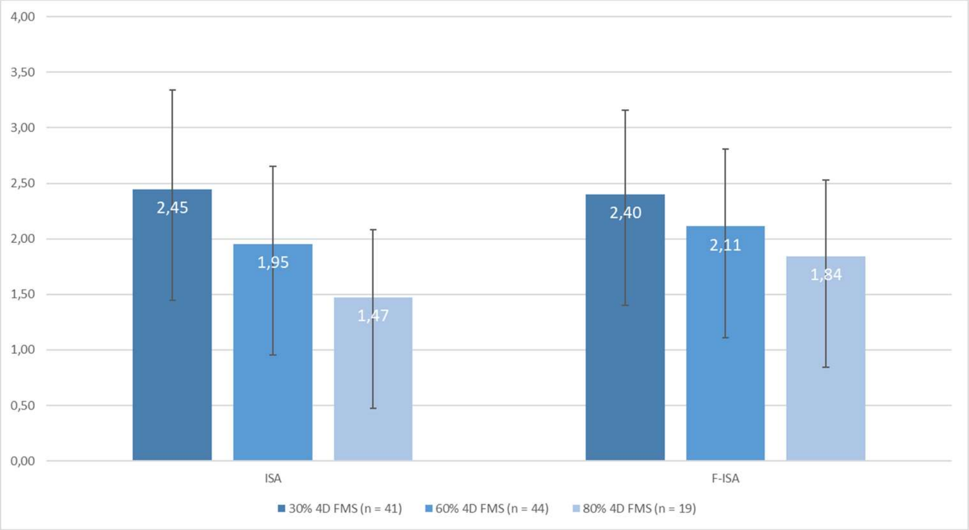


Figure 30. Mean ISA and FISA ratings in dependence of traffic distribution (30% vs. 60% vs. 80%) summarized over all participants and assessment times of both validation weeks. Error bards represent standard deviations.

In summary, it is particularly noticeable in the time dependent ISA rating of the controllers that they had alternating phases of calm and rather higher activity, especially in the 60% run (Figure 31). Due to the scenario, there were fewer manually piloted aircraft and some of them appeared in groups. If they were within planned range, the controllers sorted them into the approach flow, which resulted in a heavier workload. Subsequently, things calmed down again and the workload decreased somewhat. Towards the end of the 60% scenario, more manually guided aircraft arrived again, so that the workload rose again on average.

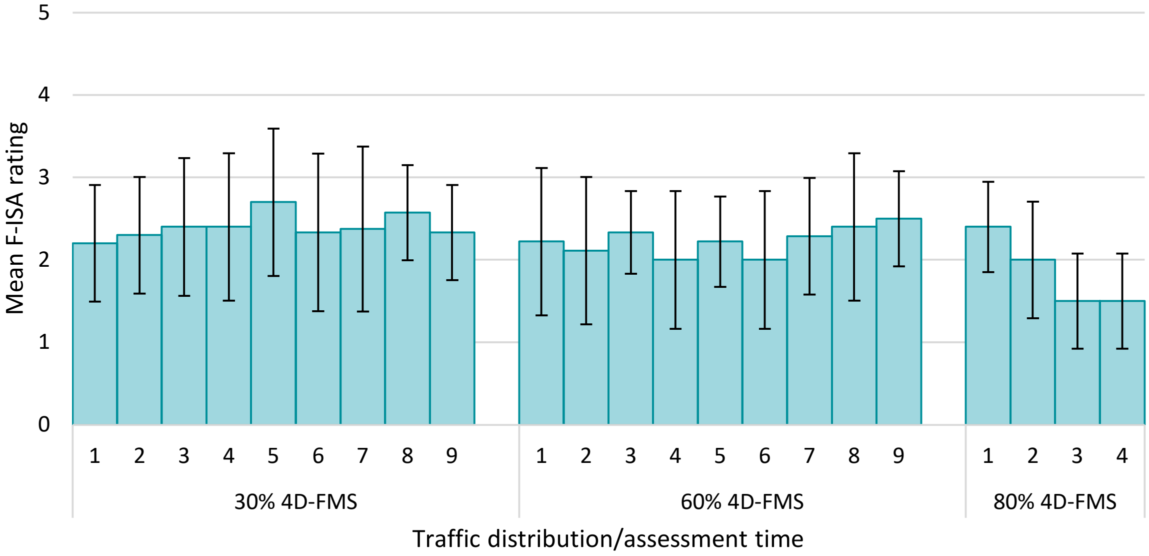


Figure 31. Mean overall ISA ratings over both trial weeks in dependence of traffic distribution (30% vs. 60% vs. 80%) and assessment time (1 – 9). Error bars represent standard deviations. Not all scenarios contain the same number of ISA/FISA queries, but the time intervals between queries are always identical.

When the RAW TLX scores of the first and second iterations are compared, it is noticeable that the controllers in the second week of the experiment felt significantly less frustrated and rated their own performance as higher. Looking at both trial weeks together, it is clear that all controllers rate all indicators lower and thus better in the 60% scenario (Figure 32). Only their own performance was seen as higher in the 30% scenario. This may be related to the fact that they were less actively involved in the scenario with more CDO aircraft.

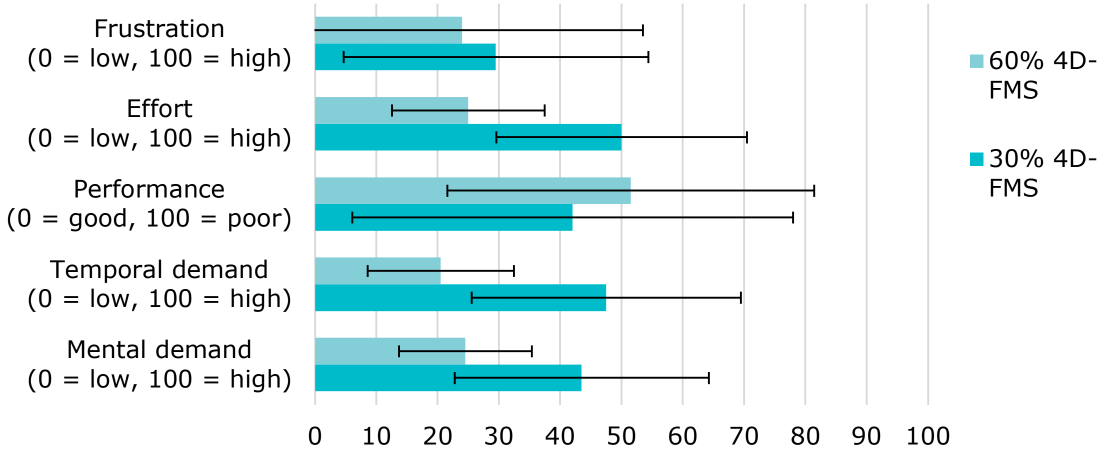


Figure 32. Mean Raw TLX scores over both trial weeks in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

4.1.4.2 SITUATION AWARENESS

Situation awareness was assessed during both validation iterations. The results are summarized in Table 18. SASHA was administered post-run and mean SASHA scores were calculated for both iterations. A higher score represents higher situation awareness and is thus desirable. Furthermore, participants rated a tailored statement regarding situation awareness in the final tailored questionnaire post-exercise. Ratings ranged between 1 (strongly disagree) to 5 (strongly agree). Mean ratings were calculated for both iterations, with a mean rating of less than 3 indicating an unacceptable level of situation awareness.

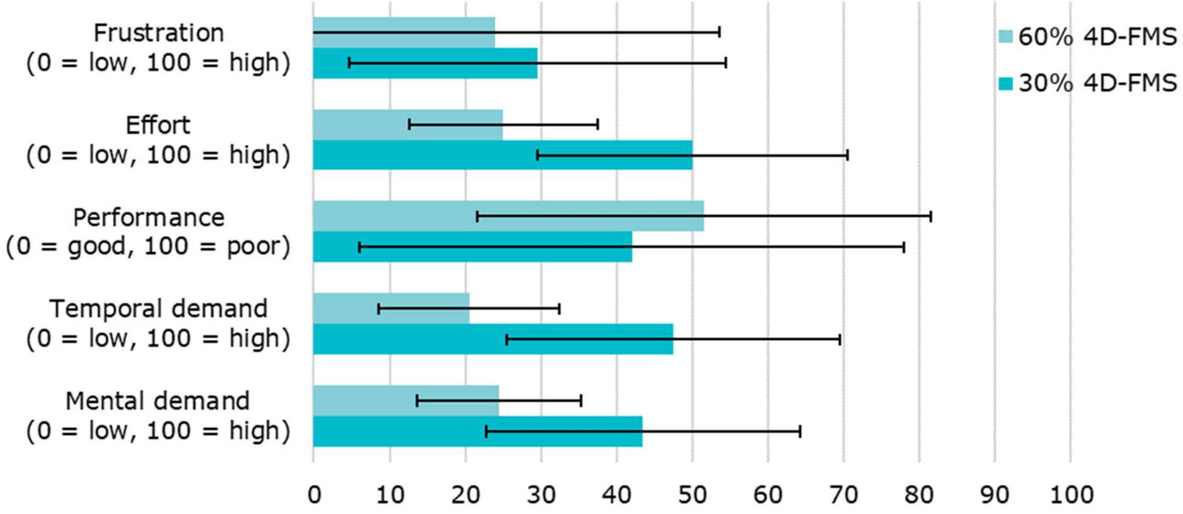


Figure 33. Mean Raw TLX scores over both trial weeks in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

Table 18. EXE-001 - Validation results for HUM – GREAT – CRT-06-10.

Criteria ID	Validation Criteria	Success	Validation Result	VALIDATION OBJECTIVE STATUS
HUM-GREAT-CRT-06-10	The level of situational awareness is within acceptable limits.		Mean SASHA scores were above the mid-point of the scale on average for all simulation runs. Overall, participants agreed to having a good mental picture of the situation. The ghosts and target windows were seen as beneficial regarding situation awareness, as reported during the debriefing. One ATCO pointed out that a too high share of 4D-FMS equipped aircraft could result in a loss of situation awareness.	OK

FIRST ITERATION

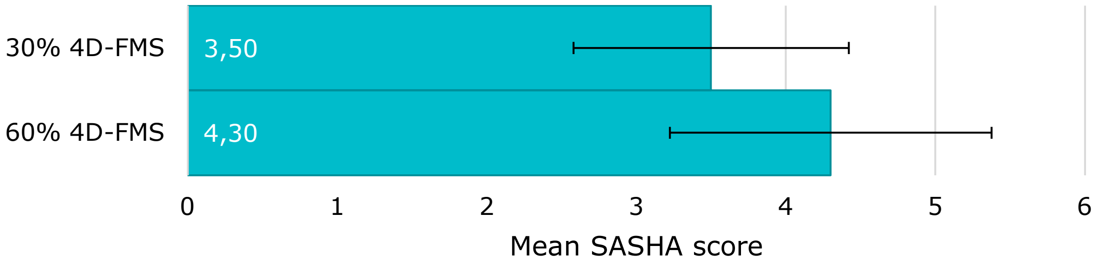


Figure 34. Mean SASHA scores in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

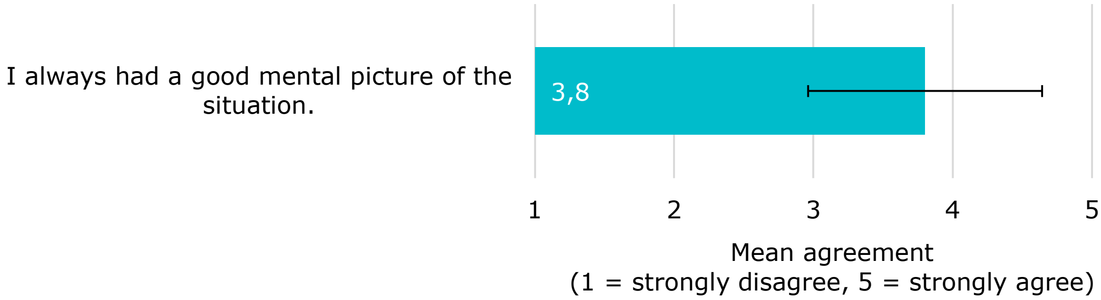


Figure 35. Mean agreement to the tailored statement regarding situation awareness. Error bars represent standard deviations.

Figure 34 shows the mean SASHA score for the 30% and the 60% traffic distribution. For both conditions, the SASHA score fell above the mid-point of the scale. In the 30% run, the SASHA score was lower than in the 60% run, indicating slightly higher situation awareness in the 60% run than in the 30% run on a descriptive level. Figure 35 illustrates participants' mean agreement to the tailored statement "I always had a good mental picture of the situation".

SECOND ITERATION

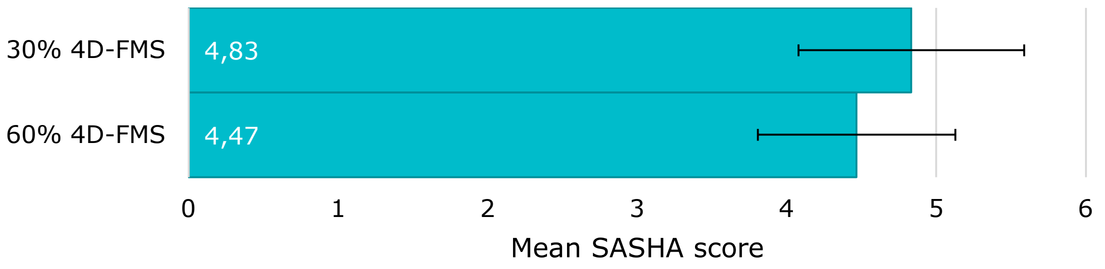


Figure 36. Mean SASHA scores in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

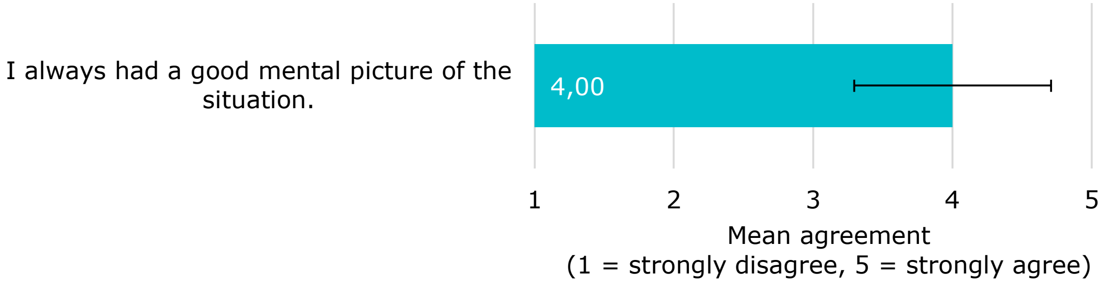


Figure 37. Mean agreement to the tailored statement regarding situation awareness. Error bars represent standard deviations.

Figure 36 depicts the mean SASHA score for the 30% and the 60% traffic distribution. For both conditions, the SASHA score fell was above 4. On a descriptive level, the SASHA scores for the 30% and the 60% condition differed only slightly, with the SASHA score in the 30% run being higher than in the 60% run. Figure 37 illustrates participants' mean agreement to the tailored statement "I always had a good mental picture of the situation". Participants' mean agreement to this was at *M*=4.00 (*SD*=0.71), indicating overall agreement with the statement.

During the debriefing, ATCOs named both the ghosts and the target windows as beneficial for increasing situation awareness. However, one ATCO raised the concern that situation awareness will be lost if the share of 4D-FMS equipped aircraft is too high.

SUM UP

If the participants in the two weeks of the experiment still showed a difference in their assessment of their situational awareness and thus an influence of the proportion of aircraft performing CDOs, this difference almost completely cancels out when viewed as a whole (Figure 38).

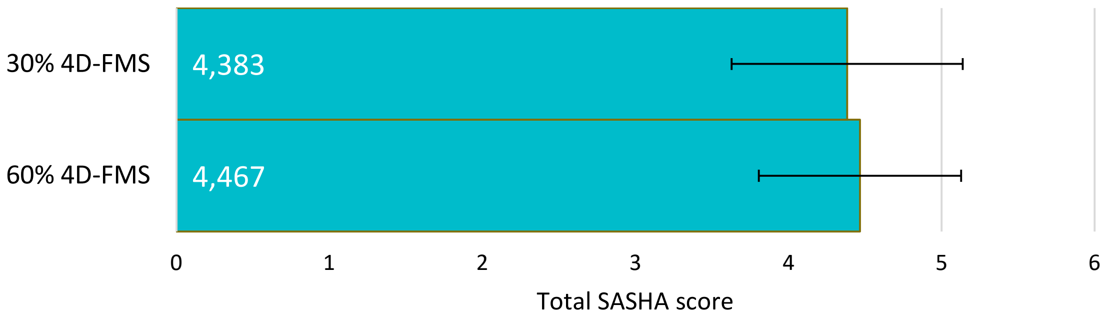


Figure 38. Mean SASHA scores over both trial weeks in dependence of traffic distribution (30% vs. 60%). Error bars represent standard deviations.

Overall, however, it must be said that the rating of just under four on the question of personal assessment of whether controllers have an overview of the overall traffic situation is not particularly high (Figure 39). This may be related to the unfamiliar controller working position, the airspace and the separation of approach flows overall, but it also shows that controllers need further dedicated support tools when using GreAT airspace.

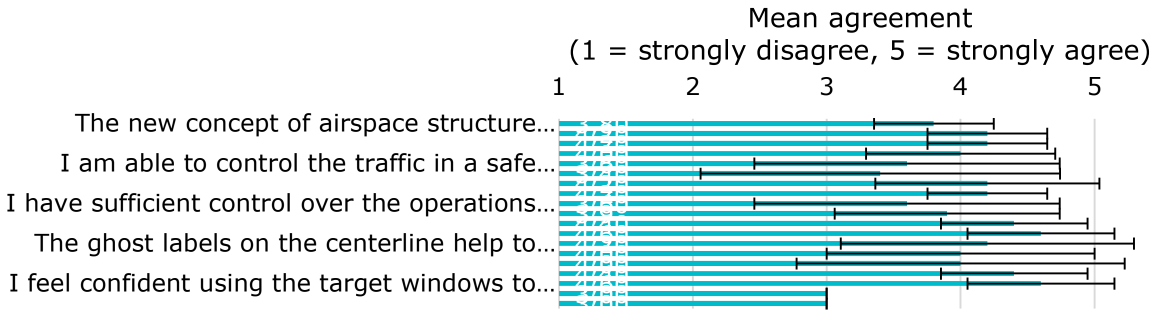


Figure 39. Mean agreement to the tailored statement regarding situation awareness over both trial weeks. Error bars represent standard deviations.

4.1.4.3 USABILITY

Usability was assessed during both validation iterations. The results are summarized in Table 19. Participants filled out the SUS post-exercise. A total SUS score of 0 represents the worst possible usability and a total SUS score of 100 the best possible usability. The mean SUS score was calculated across participants. Usability was furthermore assessed in the final tailored questionnaire post-exercise where participants were asked to give ratings between 1 (strongly disagree) to 5 (strongly agree). Mean ratings were calculated, with a mean rating of less than 3 indicating insufficient usability.

Table 19. EXE-001 - Validation results regarding usability.

Criteria ID	Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
HUM-GREAT-CRT-07-10	There is no discrepancy between system-provided information and user-required information.	In order to safely guide the aircraft through the new airspace structure, the supporting tools mainly AMAN, ghosts and Target Windows provided the required information about the projected aircraft position on the final in line with the sequence computed by the AMAN. The provided information was reported by ATCO as helpful to get the whole picture of the traffic situation and to plan ahead. Information provided by the Ghost was even considered by the ATCO as necessary safety wise. Only the information provided by the final distance indicator was not always used/ needed. Participants made also several suggestions to further improve the provided data. This indicates that there is room for improvement.	OK
HUM-GREAT-CRT-07-20	The ATCO can perform interaction without noticeable problems.	The mean SUS scores indicated "good" or close to "good" usability [Athènes 2002]. However, participants reported several challenges and proposed improvements. This indicates that there is room for improvement.	OK
HUM–GREAT–CRT-05-40	The look-and-feel of the HMI is acceptable.	All tailored statements regarding the HMI were given a mean agreement rating of 3 (neither agree nor disagree) at the minimum. However, participants proposed several improvements to the HMI, indicating that the look-and-feel of the HMI can be improved.	OK

- **FIRST ITERATION**

  - **HUM-GREAT-CRT-07-10**: As reported in the debriefing session, the ATCOs got the information they needed to do their tasks from the supporting tools and systems. As shown by Figure 40, participants rated the usefulness of the features route separation, LMP and target windows between 3 (somewhat useful) and 4 (very useful) on average. The ghosting feature was given an average rating slightly above 4 (very useful). Regarding the final distance indicator, four of five ATCOs reported during the explorative run that they did not use it.

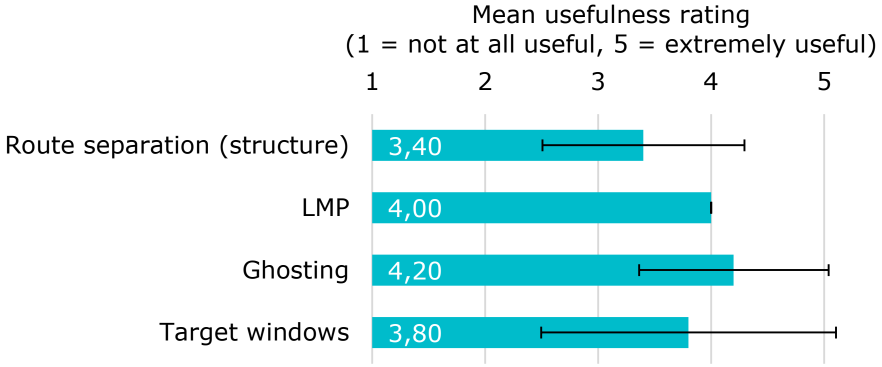


Figure 40. Mean usefulness ratings for the technical and procedural features. Error bars represent standard deviations.

Several improvements were also proposed during the debriefing and the explorative run to further complete/ adapt the provided dada to their needs. Examples of these suggestions can be found in ANNEX 7.1.1. Some of them were already implemented and integrated in the system used for the second iteration.

HUM-GREAT-CRT-07-20:

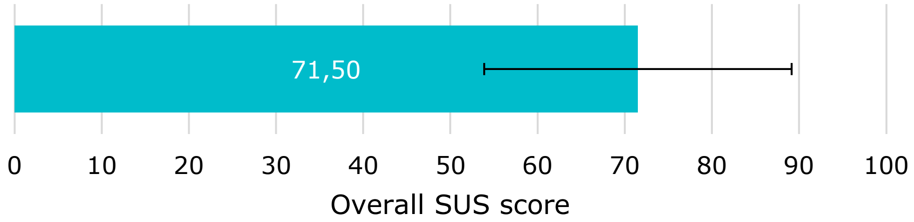


Figure 41. Mean SUS score. The error bar represents the standard deviation.

Figure 41 shows the mean SUS score. According to Bangor, a mean SUS score of 71.40 corresponds to an adjective rating “good” [Bangor 2009].

During the debriefing, all five participants reported experiencing at least one challenge. This was mostly due to the system, the airspace structure and the workflow being new. The reported challenges and proposed improvements can be found in ANNEX 7.1.2.

HUM-GREAT-CRT-05-40

Figure 42 shows the mean agreement to tailored statements regarding the system’s HMI. All statements were given an average rating of 3 (neither agree nor disagree) at the minimum. Departures and arrivals as well as 3D and 4D aircraft seem to be sufficiently distinguishable as both corresponding statements received a mean rating of at least 4 (agree). The understandability of the graphical display of the ghosts and target windows were identified as areas for improvement seeing as standard deviations for the corresponding statements were rather large and the mean agreement ratings were below 4.

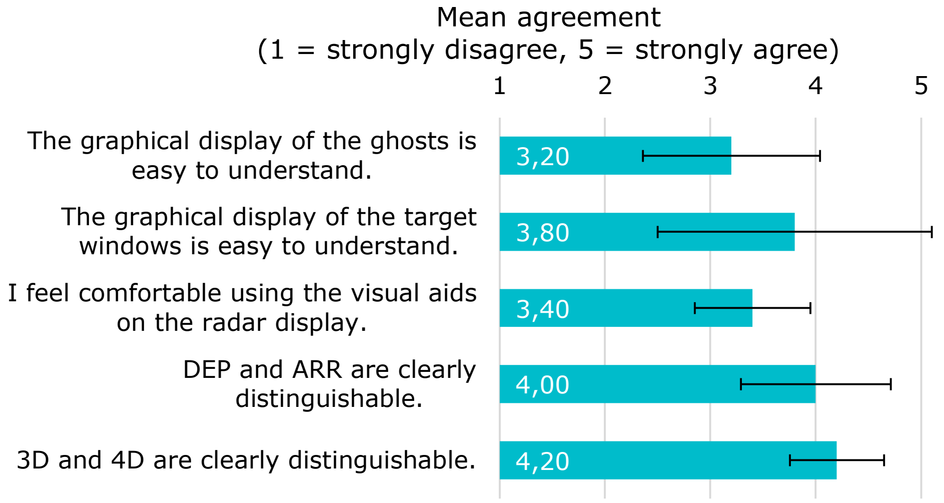


Figure 42. Mean agreement to tailored statements regarding the HMI. Error bars represent standard deviations.

During the debriefing and the explorative run, participants proposed several improvements regarding the visual design of the individual technical features. A list can be found in ANNEX 7.1.3.

SECOND ITERATION

- **HUM-GREAT-CRT-07-10**: during debriefing, controllers reported that the information provided by the supporting tools especially AMAN, the ghosts and target windows are useful and oft necessary to efficiently and safely control the traffic. The provided features made also the work easier. Only the information provided by the final distance indicator was not always used/ needed. ATCOs explained that either by the location of the distance indicator on the bottom of the screen (outside the area of attention of the ATCOs) or because they did already get the information from the displayed scale on the final. One ATCO explained also that such distance is not anymore required when the aircraft is already in the final (too late for decision making). To improve/ complete the required information, participants proposed several possibilities for improvement regarding the provided information during debriefings and the explorative run. A list can be found in ANNEX 7.1.1.

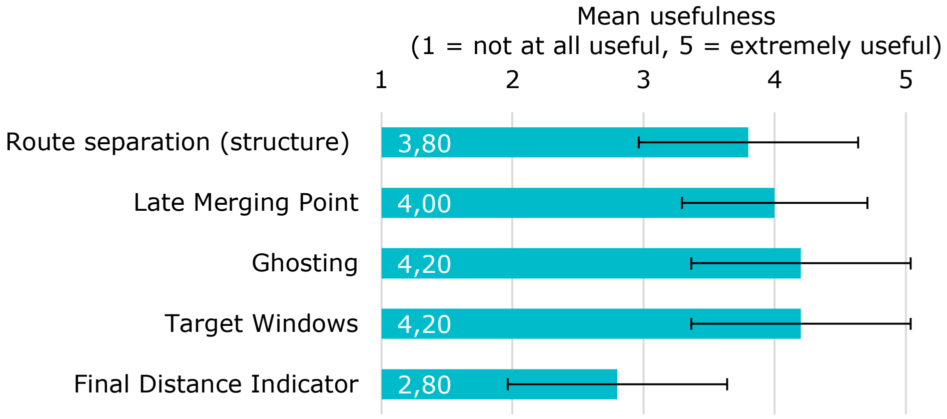


Figure 43. Mean usefulness ratings for the technical and procedural features. Error bars represent standard deviations.

Figure 43 depicts the mean usefulness ratings participants gave to the technical and procedural features. For the route separation, the LMP, the ghosting and the target windows, the mean ratings were close to 4 (very useful). The usefulness of the final distance indicator was rated the lowest for the reasons stated above. During the debriefing, all five ATCOs reported that they did not use the final distance indicator.

HUM-GREAT-CRT-07-20:

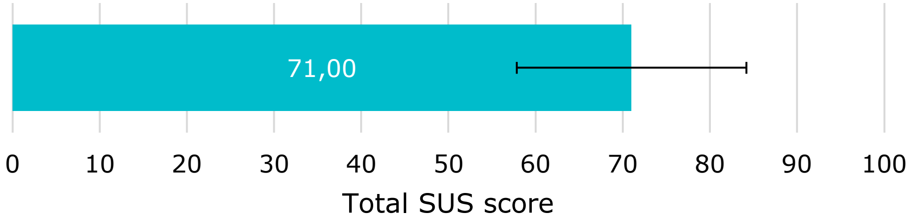


Figure 44. Mean SUS score. The error bar represents the standard deviation.

Figure 44 shows the mean SUS score. According to Bangor, Kortum and Miller [Bangor 2009], a mean SUS score of 71.40 corresponds to an adjective rating "good".

ANNEX 7.1.2 lists improvements that were proposed by participants during the runs and debriefings regarding interactions with the system.

HUM-GREAT-CRT-05-40

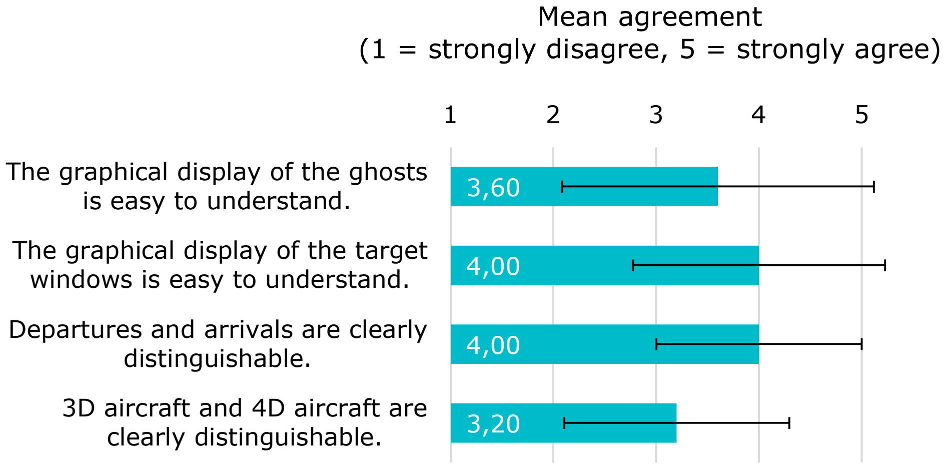


Figure 45. Mean agreement to tailored statements regarding the HMI. Error bars represent standard deviations.

Figure 45 shows the mean agreement to tailored statements regarding the system's HMI. All statements were given average ratings between 3 (neither agree nor disagree) and 4 (agree). Overall, there were large standard deviations for all statements, indicating a wide distribution of ratings.

A list of proposed improvements regarding the look-and-feel of the HMI can be found in ANNEX 7.1.3

SUM UP

The opinions of the ten test participants also differ significantly in their assessment of the usefulness of the airspace structure and support tools, but satisfaction is clearly evident with the LMP and the Ghosting support function (Figure 46). If the TargetWindows also show a broad agreement in the overall view, however, it is noticeable in the more detailed analysis that the dispersion is significantly greater: The opinions on this display thus clearly diverged among the test controllers. The Final Distance Indicator was considered to be more or less superfluous. During the debriefings, we also regularly received feedback that the controllers had not taken it into account at all when guiding the inbounds.

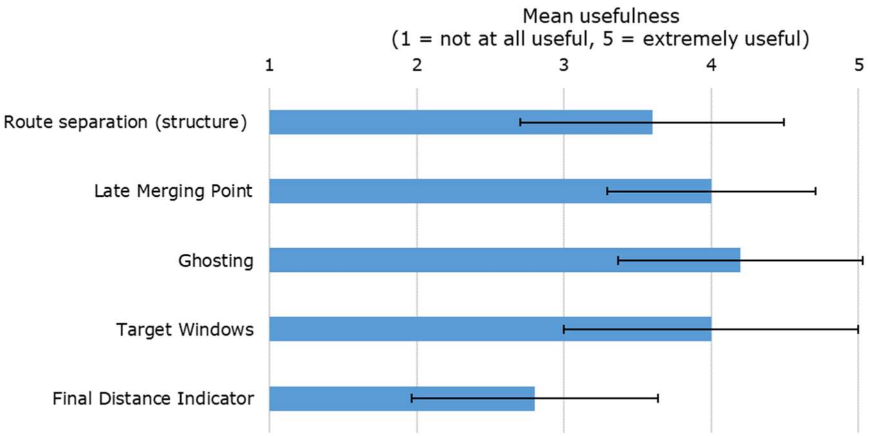


Figure 46. Mean usefulness ratings for the technical and procedural features considered over both trial weeks together. Error bars represent standard deviations.

In the SUS test for ease of use, both experimental groups consistently indicated "good" or nearly "good" ease of use. They missed a few additional approach guidance functionalities in the simulation facilities that they would have available in Budapest.

Figure 47 also shows an average level of agreement with the system's HMI when viewed overall across both experimental groups. All statements received average ratings between 3 (neither agree nor disagree) and 4 (agree). Overall, there were large standard deviations for all statements, indicating a wide distribution of ratings.

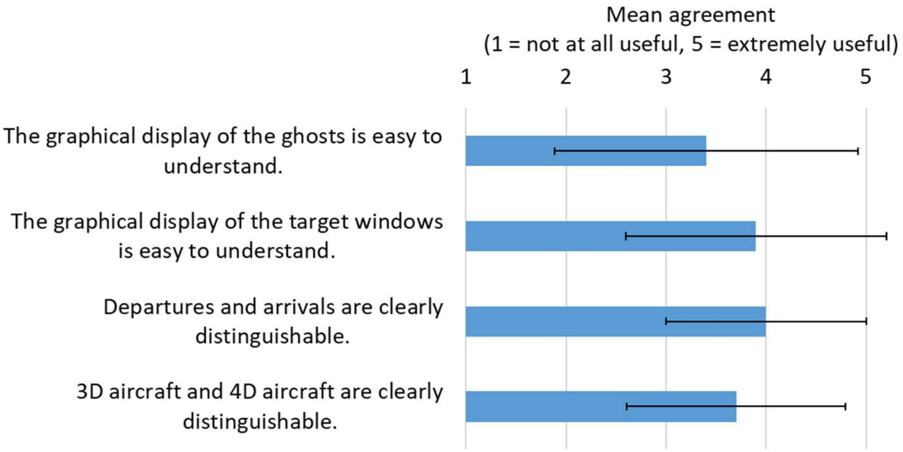


Figure 47. Mean agreement to tailored statements regarding the HMI considered over both trial weeks together. Error bars represent standard deviations.

4.1.4.4 TRUST

Trust was assessed using SATI during both iterations. The results are summarized in Table 20. The SATI was administered post-exercise and mean ratings were computed. The subscales utility, reliability, accuracy, understanding, robustness and confidence are rated from 0 (never) to 6 (always). The total SATI score is obtained by calculating the mean of the 6 subscales, with a higher score indicating a higher level of trust.

Table 20. EXE-001 - Validation results regarding HUM – GREAT – CRT-08-10.

Criteria ID	Validation Criteria	Success	Validation Result	VALIDATION OBJECTIVE STAT
HUM – GREAT – CRT-08-10	The level of trust is experienced as sufficient by the ATCO.		The SATI questionnaire indicated an overall slightly above medium level of trust into the system. The robustness of the system was identified as an area of improvement in particular. Understandability of the system was rated especially high.	OK

FIRST ITERATION

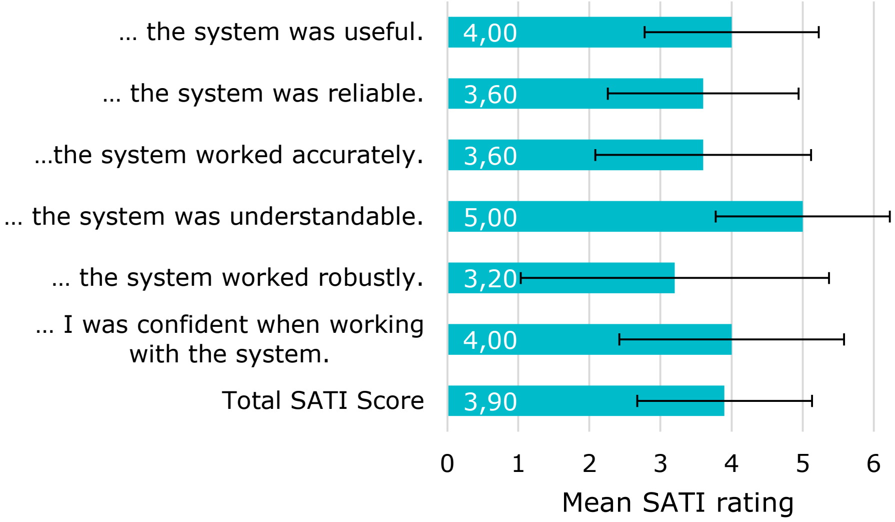


Figure 48. Mean SATI ratings. Error bars represent standard deviations.

Figure 48 shows the mean SATI scores as well as the mean total SATI score. All mean ratings were above the mid-point of the scale. Mean ratings ranged between 3 and 4 with the exception of the subscale understandability, which was given the highest mean rating of *M*=5.00 (*SD*=1.23). The subscale robustness was identified as an area for improvement as it was given the lowest mean rating with a noticeably high standard deviation. One possible explanation for this could be the fact that several technical issues occurred during the simulations in both validation iterations. Overall, these results were interpreted as a sufficient level of trust.

SECOND ITERATION

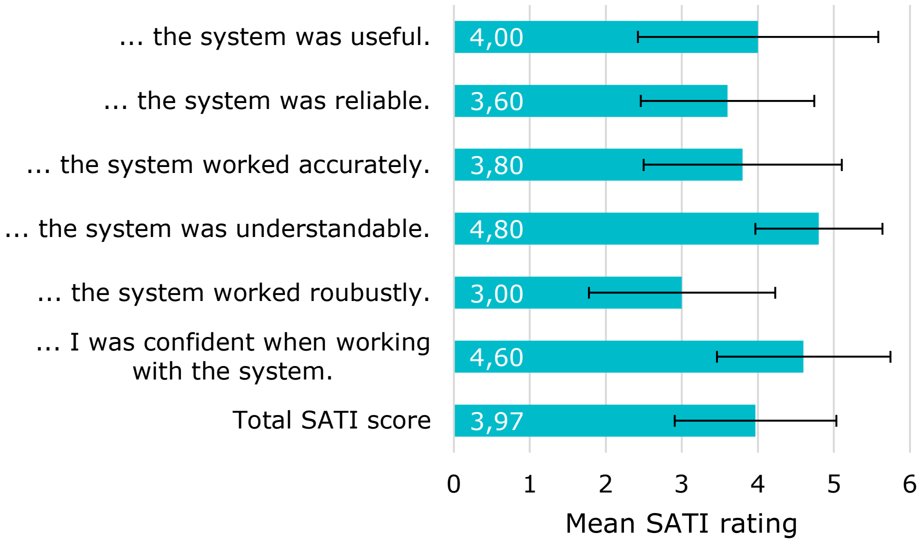


Figure 49. Mean SATI ratings. Error bars represent standard deviations.

Figure 49 shows the mean SATI scores as well as the mean total SATI score. All mean ratings were at the mid-point of the scale or above. In accordance with the first iteration, the subscale understandability was given the highest mean rating while the subscale robustness was given the lowest mean rating. One possible explanation for this could be the fact that several technical issues occurred during the simulations in both validation iterations. Overall, these results were interpreted as a sufficient level of trust.

4.1.5. CAPACITY

Although an increase of capacity was not targeted by this solution, during the validation activities and subsequent analysis of the data, it has been observed that ATC assistance tools have effectively supported them in guiding the traffic. This situation is reflected in the results presented below (Figure 50 displays a comparison between number of approaches executed in two various scenarios where different distribution of 3D-FMS and 4D-FMS operations has been analysed. The blue bars represent the situation where 30% of the approaching traffic distribution was attributed to the flights with FMS equipment, while the green bars correspond to 60% of such equipment. Taking that into consideration, it can be easily pointed out that a greater number of aircraft were more efficiently routed for landing by all ATCs.

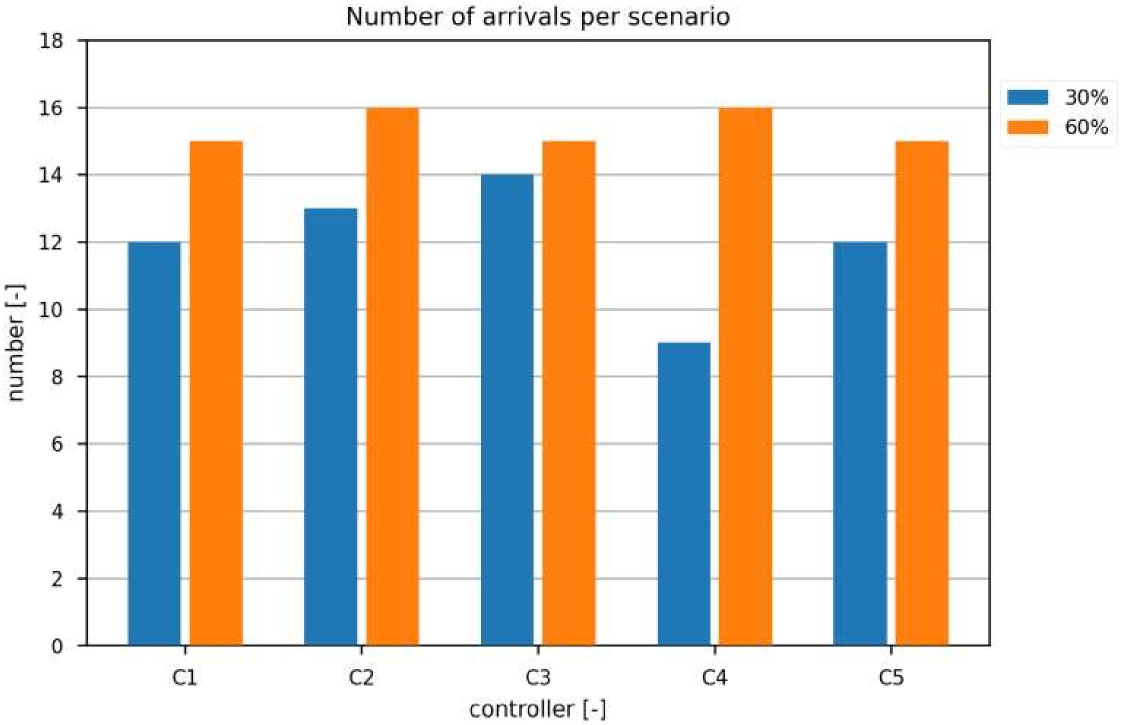


Figure 50. Capacity assessment from validation trials.

4.1.6. FEASIBILITY AND ACCEPTABILITY OF CONCEPT

This chapter contains results regarding the feasibility and acceptability of the concept that cannot be mapped to any of the objectives. This includes tailored questions that were administered in the final tailored questionnaire post-exercise as well as answers to debriefing questions. The tailored questions were rated from 1 (strongly disagree) to 5 (strongly agree) and mean ratings were calculated. Mean ratings of 3 or higher were interpreted as general agreement to the statements.

FIRST ITERATION

Figure 51 shows the mean agreement ratings to tailored questions regarding the overall concept. Participants rated most statements between 3 (neither agree nor disagree) and 4 (agree) on average.

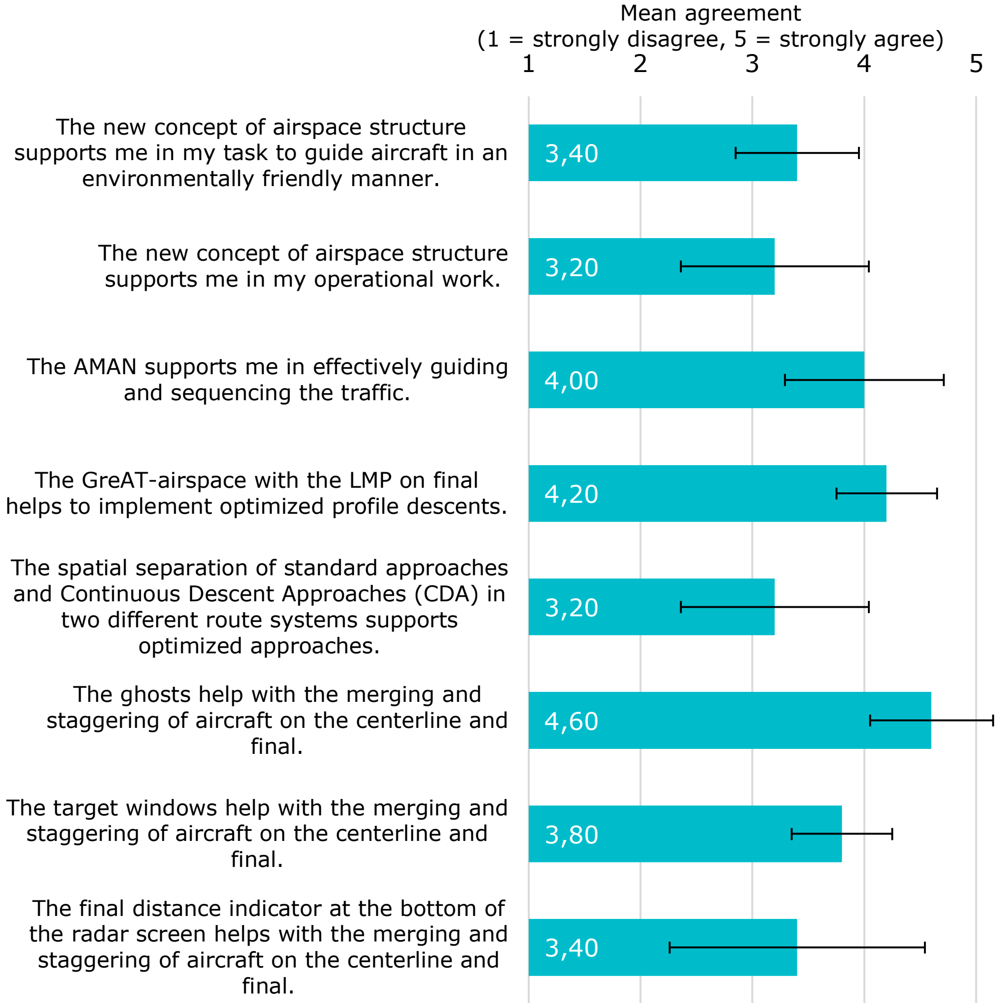


Figure 51. Tailored statements regarding the overall concept. Error bars represent standard deviations. Error bars represent standard deviations.

During the debriefing, participants were asked further questions regarding the overall concept. The questions and answers are summarized in Table 21.

Table 21. Debriefing questions from the first iteration regarding the overall concept.

Debriefing questions
Which features of the coupling of AMAN and FMS functionalities in TMA provide benefits and in what way?
• **This question was answered by all five ATCOs** • **All five participants stated that they found the system generally helpful** • **Benefits mentioned were:** ○ **The system helps to manage high traffic load** ○ **The system is beneficial for route separation** ○ **The system helps to make predictions** ○ **One participant stated that the system could open a new way of controlling** • **However, participants also raised some criticism:**

  - ○ **One participant stated that the system is not efficient for low traffic events and suggested a T-bar-structure as a solution**

Which features/ supporting tools would you like to see implemented in real ATC?

- **This question was answered by all five ATCOs**
- **Three participants would like to see both the ghosts and the target windows implemented**
- **One participant stated that he would like to see the ghosts implemented**
- **One participant named the target windows only**

In your opinion, for which traffic distribution could the system be most helpful?

- **This question was answered by all five ATCOs**
- **All five participants judged that the system would be most helpful for a higher proportion of 4D traffic.**

Would the proposed support tools help to reduce the negative effects of merging CDAs and standard approaches?

- **This question was answered by all five ATCOs**
- **All five participants agreed that the system would help with this**

Does the LMP have the correct position to safely and efficiently merge the separated arrival streams?

- **This question was answered by all five ATCOs**
- **Two participants did not find this relevant**
- **Two participants would prefer more than 6 NM**
- **One participant found the position appropriate but suggested that 6 NM could be too close for heavy aircraft**

Do you have any additional comments, thoughts or new ideas?

- **This question was answered by two ATCOs**
- **One participant commented that it should also be possible for 3D aircraft to use the LMP and that CDA should also be possible for high traffic. For this, a sectorization of the airspace would be necessary.**
- **One participant voiced general approval of the concept.**

SECOND ITERATION

Figure 52 shows the mean agreement ratings to tailored questions regarding the overall concept. Most statements were given an average agreement around 4 (agree). The statement "The final distance indicator at the bottom of the radar screen helps with the merging and staggering of aircraft on the centreline and final" was rated with 3 (neither agree nor disagree) unanimously, indicating that the final distance indicator was not perceived as particularly helpful for the merging and staggering of aircraft.

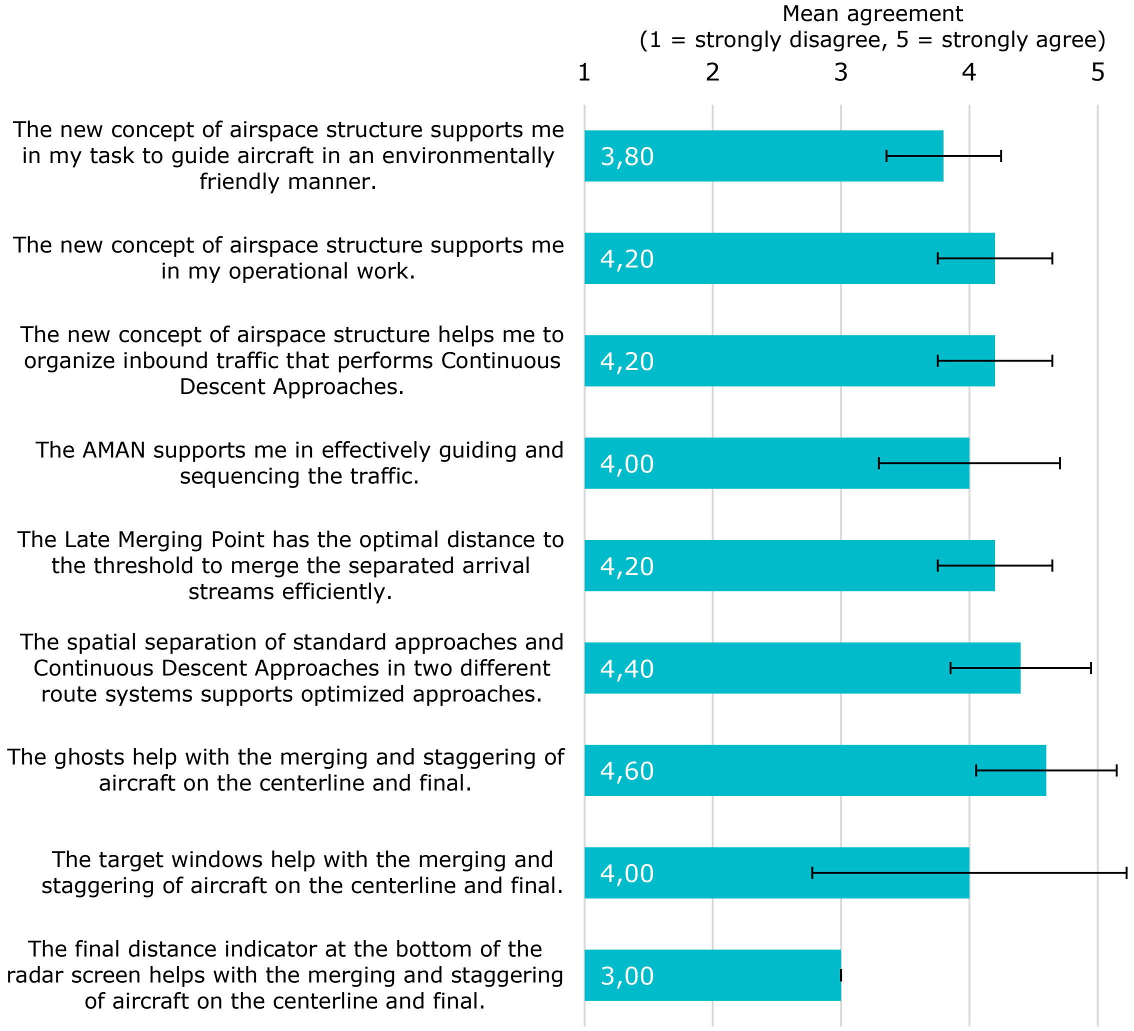


Figure 52. Tailored statements regarding the overall concept. Error bars represent standard deviations. Error bars represent standard deviations.

During the debriefing and the explorative run, participants were asked questions about the overall concept. The following feedback was given:

APPROVING FEEDBACK ABOUT THE CONCEPT

- Three ATCOs saw the system as feasible overall
- One ATCO saw "huge potential" in the route system and thought it could be beneficial for noise reduction as well.

CRITICAL FEEDBACK ABOUT THE CONCEPT

- It was criticized that the system put 3D aircraft at a disadvantage.
- According to one ATCO, shorter/more efficient routes than proposed by the system would have been possible for the 3D aircraft.
- Two ATCOs reported that conflict resolution is more difficult because the 4D aircraft are untouchables. One ATCO stated that he dislikes the idea of untouchable 4D equipped aircraft According to this ATCO, the more aircraft he can control, the better. According to the other ATCO, the concept reduces tolerance for mistakes: If a mistake is made, the ATCO has to "punish" the 3D aircraft because the 4D are untouchables. Even though this kind of conflict resolution is less efficient compared to today's

operations, the ATCO deemed the concept as more efficient overall. Both ATCOs expressed that they would like to be able to influence 4D aircraft.

- One ATCO reported that the design could impact his work by including more monitoring than controlling tasks. This was rated neither positively nor negatively.

PROPOSED IMPROVEMENTS OF THE CONCEPT INCLUDED:

- The area around the LMP was experienced as too crowded by one ATCO; there should be less routes coming from the south.
- One ATCO proposed to add predefined points on the downwind and final in order to provide clearances "direct to point"
- Additional tools may be needed to guide the traffic more efficiently
- Regarding the **ghosts**, ATCOs mentioned several benefits:
    - They help to calculate the spacing
    - They help to negotiate the time
    - They improve efficiency
    - Two ATCOs stated that the ghosts could also be interesting for 3D aircraft or other purposes than the GreAT airspace.
- Regarding the **target windows**, there were some critical comments:
    - Four ATCOs criticized that the target windows did not provide the most efficient solution.
    - Three ATCOs proposed that target windows should be smaller to improve efficiency.

Table 22 lists further questions and ATCOs’ answers regarding the overall concept.

Table 22. Debriefing questions from the second iteration regarding the overall concept.

Debriefing questions
In your opinion, for which traffic distribution (e.g. 30% 4D or 60% 4D) could the system be most helpful?
This was answered by all five ATCOs. Five differing opinions emerged: (1) **The system is useful for 50% traffic and above** (2) **The share of 4D equipped aircraft should not be higher than 40-50%** (3) **The system is always helpful** (4) **The more aircraft are under the control of the ATCO, the better. The ATCO did not like the idea of untouchable 4D aircraft. But in high traffic scenarios, the technical features are essential.** (5) **60% 4D was easier to control than 30% 4D as the ATCO felt like he had a higher capacity then**
Which features would you like to see implemented in real ATC?
This question was answered by all five participants • **The ghosts were mentioned five times** • **The target windows were mentioned four times** • **One participant would like to see the procedures implemented**

4.1.7. SUMMARY OF EXERCISES RESULTS

4.1.7.1 SUMMARY RELATED TO FEEDBACK ON CONCEPT

The concept was generally evaluated as feasible and accepted by the ATCOs. Nevertheless, the ATCOs voiced criticism about the concept. One point of criticism was the concept of untouchable 4D-FMS equipped aircraft, because (1) ATCOS felt that 3D-FMS equipped aircraft were put at a disadvantage by this and (2) efficiency of conflict resolution was impaired. Regarding efficiency, ATCOs furthermore reported that the target windows did not help to find the most efficient solution capacity wise. Still, several ATCOs reported that they would like to see the ghosts as well as the target windows implemented in real operations.

4.1.7.2 SUMMARY PER OBJECTIVE ID

The table below summarizes the exercise results for each success criteria as defined in the Validation Plan. One of the following status has been attributed to each success criteria depending on the exercise outcome:

- OK: Validation objective achieves the expectations (exercise results achieve success criteria)
- NOK: Validation objective does not achieve the expectations (exercise results do not achieve success criteria)
- Not Addressed: Validation objective could not be analysed (mostly when the event to consider did not occur or when no data was recorded)

Table 23. EXE-001 - Validation Criteria status

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	VALIDATION & OBJECTIVE STATUS
ENVIRONMENT				
ENV – GREAT – OBJ-01	To assess the reduction of exhaust emissions due to solution	ENV – GREAT – CRT-01-10	The solution results in reduction of exhaust emissions compared to the reference scenario.	*Still under progress / to be included in the VALR final iteration*
OPERATIONAL EFFICIENCY				
OPE – GREAT - 02	To assess the reduction in flown distance per aircraft due to solution	OPE – GREAT – CRT-02-10	The distance flown is reduced compared to reference scenario.	OK

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	VALIDATION & OBJECTIVE STATUS
OPE – GREAT - 03	To assess reduction in fuel burn due to solution	OPE – GREAT – CRT-03-10	The average fuel burn by aircraft is reduced compared to the reference scenario.	Combined with ENV – GREAT – OBJ-01
CAPACITY				
CAP – GREAT - 04	To assess the solution’s impact on capacity	CAP – GREAT – CRT-04-10	The solution does not reduce capacity.	OK
HUMAN PERFORMANCE – WORKLOAD				
HUM – GREAT - 05	To assess the ATCO’s workload	HUM – GREAT – CRT-05-10	The level of workload is within acceptable limits.	OK
HUMAN PERFORMANCE – SITUATIONAL AWARENESS				
HUM – GREAT - 06	To assess the ATCO’s situational awareness	HUM – GREAT – CRT-06-10	The level of situational awareness is within acceptable limits.	OK
HUMAN PERFORMANCE – USABILITY				
HUM – GREAT - 07	To assess the usability of the system	HUM-GREAT-CRT-07-10	There is no discrepancy between system-provided information and user-required information.	OK
		HUM-GREAT-CRT-07-20	The ATCO can perform interaction without noticeable problems.	OK
		HUM – GREAT – CRT-05-40	The look-and-feel of the HMI is acceptable.	OK
HUMAN PERFORMANCE – TRUST				
HUM – GREAT - 08	To assess the ATCO’s trust in the system	HUM – GREAT – CRT-08-10	The level of trust is experienced as sufficient by the ATCO.	OK
SAFETY				
SAF – GREAT - 09	To assess the impact on the safety level of the system	SAF – GREAT – CRT-09-10	Procedures and system functions are safe in normal situations.	OK
		SAF – GREAT – CRT-09-20	Procedures and system functions are safe in abnormal situations.	N/A

4.2. VALIDATION EXERCISE EXE-002 – HC AND PILDO LABS

The results are sorted by the KPAs environment, operational efficiency and safety, and under each KPA by the type of validation (RTS and Shadow mode).

4.2.1. ENVIRONMENT

In this KPA, it is assessed whether the use of MergeStrip will have a positive impact on the environment.

Table 24. EXE-002 - Environment KPA results.

Objective Criteria ID	Exercise Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
ENV–GREAT-01	Less CO_2 emitted compared to reference scenario	It was observed with the use of DailyFuel application that the use of MergeStrip brings minor improvements in the mean values of all the analysed indicators: fuel consumption and CO_2 emissions.	OK

4.2.1.1 AVERAGE FUEL BURN PER FLIGHT

The average fuel burn per flight has been computed with the tool DailyFuel, developed by Pildo Labs. The results are summarized in Table 25.

Table 25. EXE-002 – Fuel consumption results.

Indicator	MergeStrip	No MergeStrip
Fuel consumption	361.26 kg	364.24 kg

4.2.1.2 AVERAGE CO2 EMISSIONS PER FLIGHT

The average CO2 emissions per flight have been computed with the tool DailyFuel, developed by Pildo Labs. The results are summarized in Table 26.

Table 26. EXE-002 – CO_2 emissions results.

Indicator	MergeStrip	No MergeStrip
CO_2 emissions	1137.93 kg	1147.36 kg

4.2.2. SAFETY

4.2.2.1 SAFETY PERFORMANCE

Safety performance was evaluated separately for real time simulations (RTS) and shadow mode trials.

4.2.2.1.1SAFETY PERFORMANCE IN RTS

Table 27. EXE-002- Safety Performance Results.

Objective Criteria ID	Exercise Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
SAF–GREAT-09	According to ATCOs the predictions of the "what if function" was adequate for safe service provision	All the participating ATCOs agreed that the what-if function did not decrease the level of safety.	OK
SAF–GREAT-09	The working of "what if function" was appropriate	50% of the ATCOs gave positive response to the acceptability of the what-if predictions from safety point of view. There are still bugs to correct (e.g. aircraft moving backwards) and improvements to be made, which are outlined in this section.	POK
SAF–GREAT-09	The working of "what if function" was appropriate	The scenarios haven't been specifically targeting abnormal scenarios, however, there were two instances when unexpected situations occurred. One of those prompted the runway change in fact. The ATCOs had no negative experiences working with the what-if function when there was an unexpected situation, however, the runway change functionality has to be further improved.	POK

Safety related questions have been integrated into the post-simulation questionnaire. Figure 53 shows the results, separated into the two iterations. The charts indicate that the majority of ATCOs thought that the what-if function did not decrease the level of safety, however, the predictions were not completely acceptable. There were various reasons behind this experience: for instance, the reference point was not in line with the ATCOs' expectations, and some of the aircraft were incorrectly appearing again after they had landed and were moving backwards on the MergeStrip line.

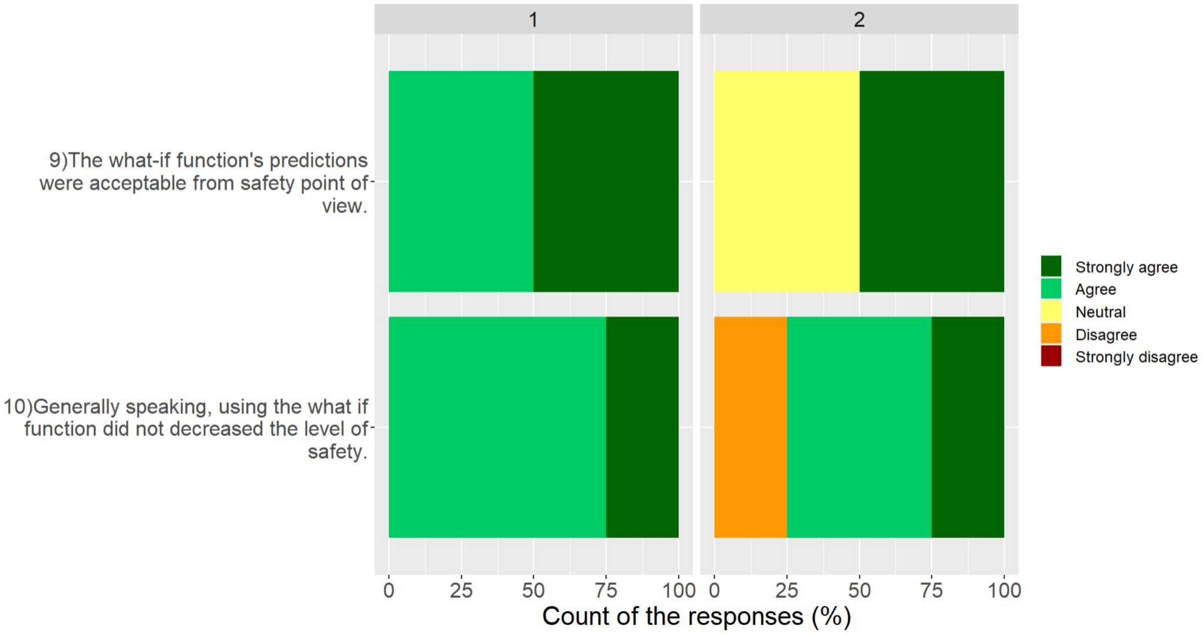


Figure 53. Safety related question results from the post-simulation questionnaire.

No specific scenarios have been created for abnormal or degraded modes, however, there was one occasion when due to medical emergency the runway direction had to be changed. This enabled the ATCOs to test the runway change functionality and give feedback on how to improve it (see Table 31 and Table 33 for details).

4.2.2.1.2SAFETY PERFORMANCE IN SHADOW MODE

Table 3. EXE-002- Safety Performance Results.

Objective	Success criteria	Result	Status
SAFETY			
To assess safety of the logic behind system functions in normal situations	According to ATCOs the punctuality of "ETA prediction function" was adequate for safe service provision	Mostly positive, but concerns have been raised.	NOK
	According to ATCOs the predictions of the "what if function" was adequate for safe service provision	Mostly negative feedbacks have been received.	NOK
	According to ATCOs the logic behind conflict resolution advisory was reasonable and adequate for safe service provision (HF-TRUST)	The recommended resolution was often considered inadequate.	NOK
To assess safety of system functions in normal situations	The working of "ETA prediction function" was appropriate	Mostly negative feedbacks have been received.	NOK

	The working of "what if function" was appropriate	Mixed reviews were received, overall below the success criterion.	NOK
	The working of conflict resolution advisory was appropriate	The recommendation was often considered inadequate, the ATCOs reported threats to safe operation and work.	NOK
	The number of separation minima infringements is not higher	The unreliable and sometimes broken functioning of the system made it difficult to make a proper assessment.	Not applicable, could not be properly tested.
To assess safety of system functions in abnormal situations	The working of "ETA prediction function" was appropriate	There was a wide variation in responses, but overall the result is below the acceptance threshold.	NOK
	The working of "what if function" was appropriate	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
	The working of conflict resolution advisory was appropriate	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
To assess safety of degraded modes of system functions.	The working of fail-safe operation of "ETA prediction function" is appropriate in case of total/partial loss or corruption of function.	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
	The working of fail-safe operation of "what if function" is appropriate in case of total/partial loss or corruption of function.	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
	The working of fail-safe operation of conflict resolution advisory is appropriate in case of total/partial loss or corruption of function.	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
	The alert in case of degradation of "ETA prediction function" was useful.	The unreliable and sometimes broken functioning of the system made it	Not applicable, could not be properly tested.

		impossible to make a proper assessment.	
	The alert in case of degradation of "what if function" was useful.	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.
	The alert in case of degradation of conflict resolution advisory was useful.	The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment.	Not applicable, could not be properly tested.

Objective: To assess safety of the logic behind system functions in normal situations.

Success criteria: According to ATCOs the punctuality of "ETA prediction function" was adequate for safe service provision

The unreliable and sometimes broken functioning of the system made it difficult to make a proper assessment. However, in certain periods it was possible to draw conclusions about the function of the feature, which are reflected in the following results. Even disregarding systemic errors, some ATCOs considered the ETA calculation to be inaccurate, while others found it to be a reasonable prediction. According to an ATCO, the estimation of speed dynamics are seemed rather unrealistic (Figure 54). On the other end, some ATCOs found that the 3.0 version provided similar results than the current version and the main system. Over the shoulder observation confirmed that even when the stability issue was solved, ATCOs tended to be overcritical to the functioning of the system (i.e. even when comparing ETA of the current version and 3.0).

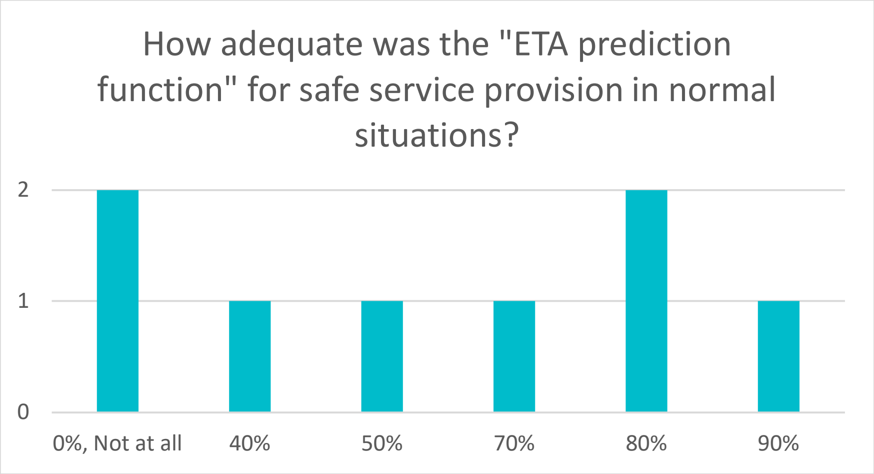


Figure 54. Safety of the logic feedback on ETA prediction in normal situation.

Objective: To assess safety of the logic behind system functions in normal situations.

Success criteria: According to ATCOs the predictions of the "what if function" was adequate for safe service provision

The assessment of the function from a safety point of view has divided ATCOs, mostly negative feedbacks have been received (Figure 55). During over the shoulder discussions certain ATCOs found it easy to interact with What-if. Those who had discomfort with e.g. the layout of the pop-up windows, tended to be more critical and patient. At the other end of the spectre, certain ATCOs were satisfied with the concept and its implementation.

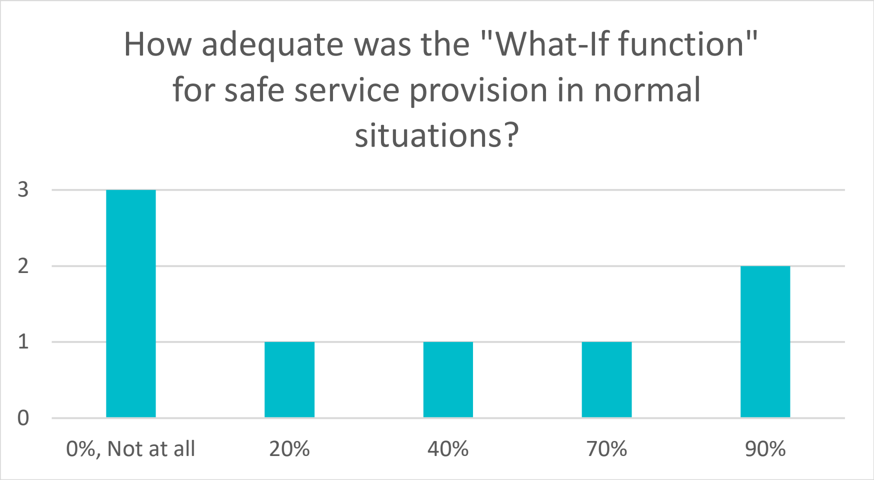


Figure 55. Safety of the logic feedback on "'What-if' function" in normal situation.

Objective: To assess safety of the logic behind system functions in normal situations.

Success criteria: According to ATCOs the logic behind conflict resolution advisory was reasonable and adequate for safe service provision (HF-TRUST)

During debriefing, the ATCOs reported negative experiences where in similar situations the recommender gave different advice for similar situations. The advice was often considered inadequate (Figure 56).

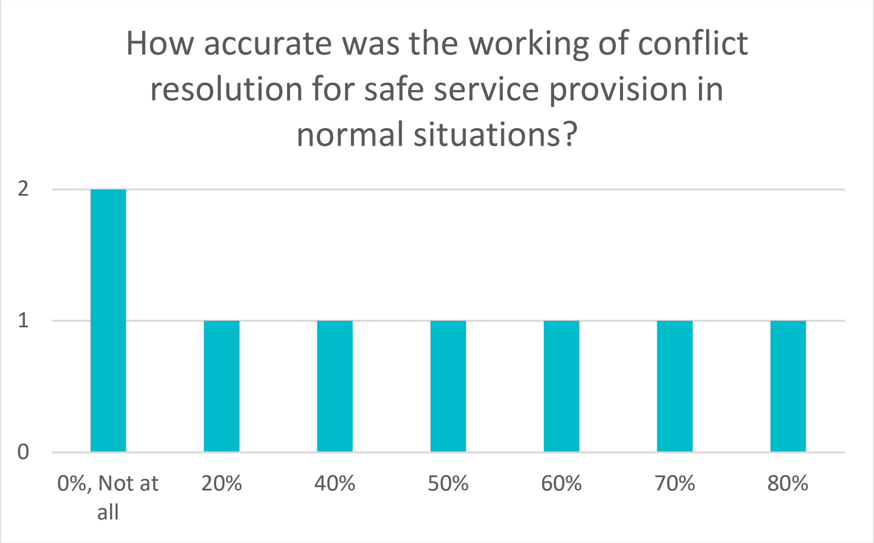


Figure 56. Safety of the logic feedback on conflict resolution in normal situation.

Objective: To assess safety of system functions in normal situations.

Success criteria: The working of "ETA prediction function" was appropriate

Feedbacks have been mixed, but rather negative about safe operation of "ETA prediction function" in practice. No one found it truly satisfactory, no 90% or 100% response received (Figure 57). Contrary to opinion expressed by ATCOs, when not in the testing periods, consultations with PCs showed that the ETA provided similar ETAs as the version currently used in the OPS room. It must also be mentioned that 3.0 data refreshing frequency differs from that of the current version, which caused an annoying effect on ATCOs, hence degraded the evaluation they gave.

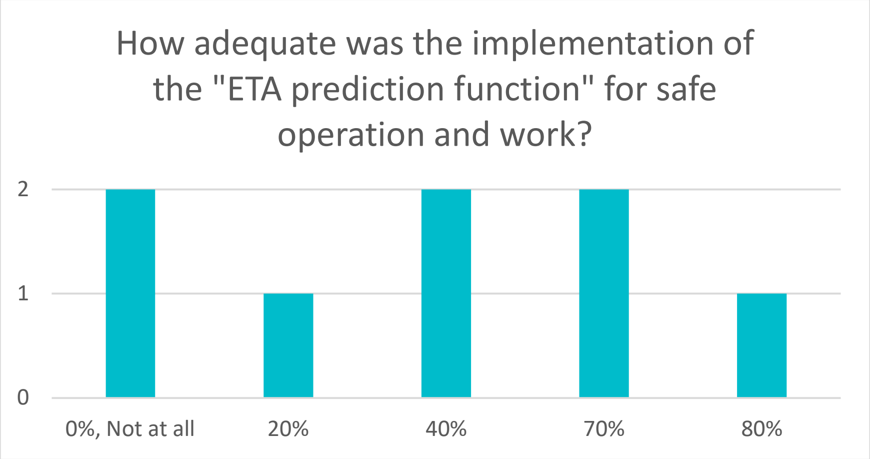


Figure 57. Safety feedback on ETA prediction in normal situation.

Objective: To assess safety of system functions in normal situations.

Success criteria: The working of "what if function" was appropriate

The implementation of "'What-if' function" was found to be adequate by ATCOs, but due to the uncertain operation, the overall perception of safety was mixed and insufficient to meet the success criteria (Figure 58). For over the shoulder observations, please see above.

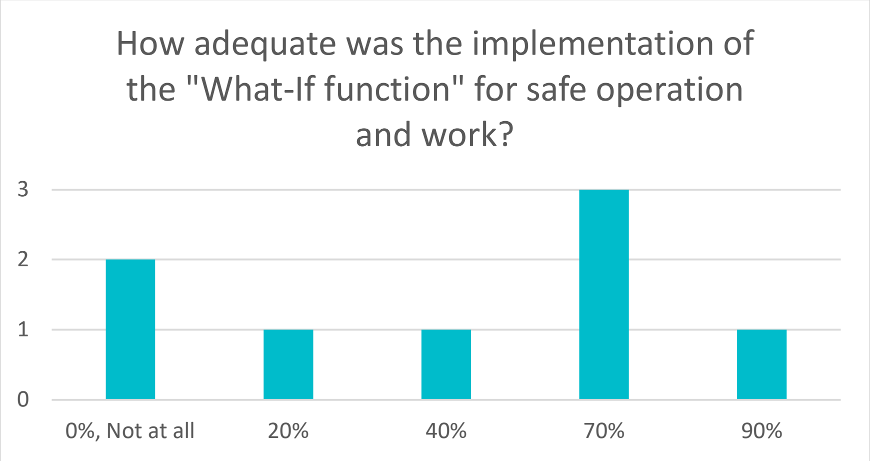


Figure 58. Safety feedback on "'What-if' function" in normal situation.

Objective: To assess safety of system functions in normal situations.

Success criteria: The working of conflict resolution advisory was appropriate

The recommendation was often considered inadequate, the ATCOs reported threats to safe operation and work (Figure 59).

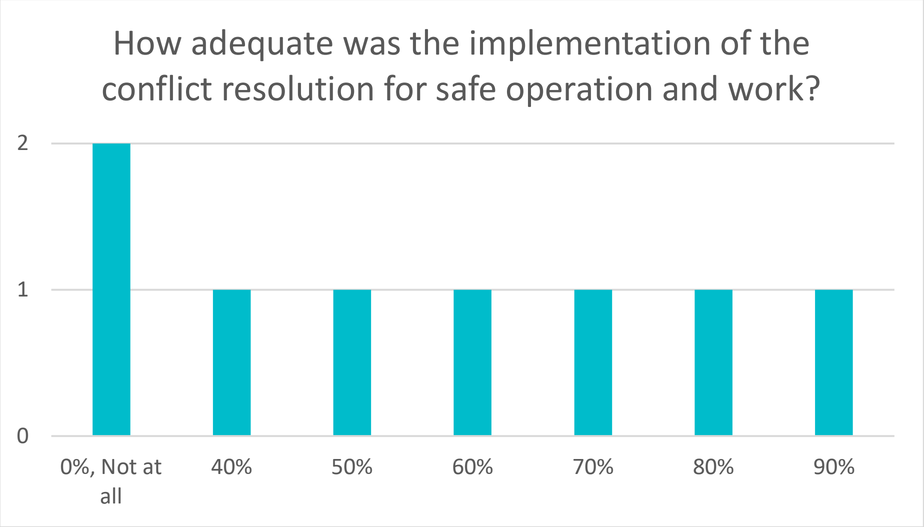


Figure 59. Safety feedback on conflict resolution in normal situation.

Objective: To assess safety of system functions in normal situations.

Success criteria: The number of separation minima infringements is not higher

The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment (Figure 60).

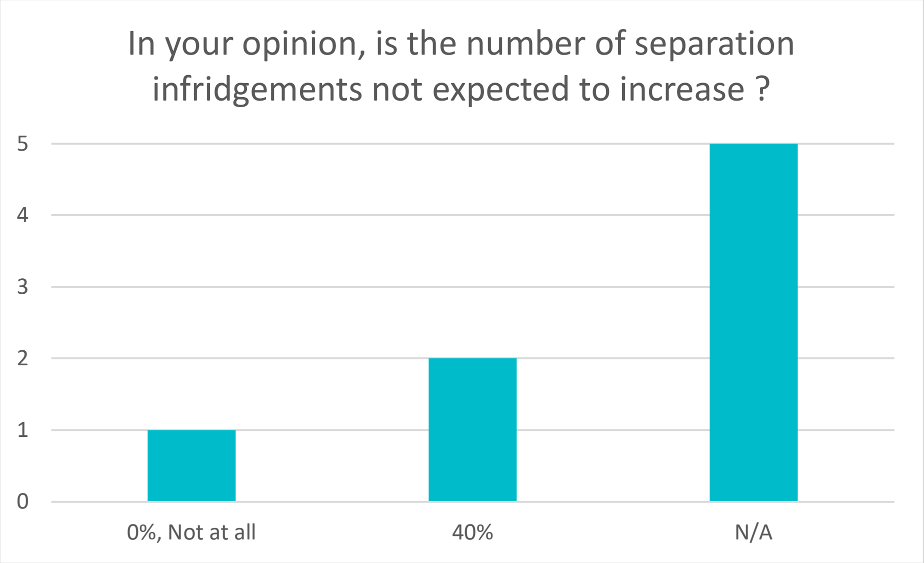


Figure 60. Safety feedback on possible separation infringements.

Objective: To assess safety of system functions in abnormal situations.

Success criteria: The working of "ETA prediction function" was appropriate.

There was a wide variation in responses, but overall the result is below the acceptance threshold. The ATCOs reported difficulties in controllability because it was not possible to follow the aircraft to landing due to a malfunction (Figure 61).

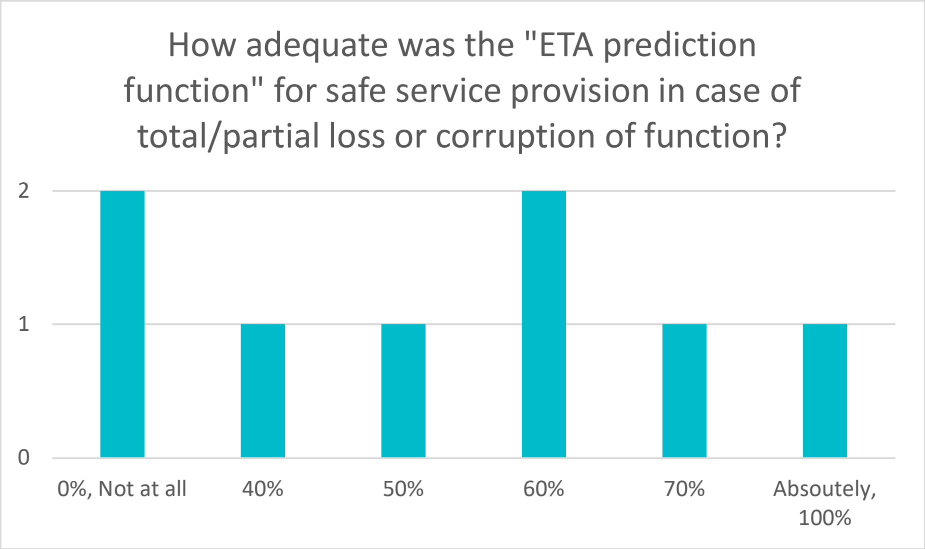


Figure 61. Safety of the logic feedback on ETA prediction in abnormal situation.

Objective: To assess safety of system functions in abnormal situations.

Success criteria: The working of "what if function" was appropriate

The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment (Figure 62).

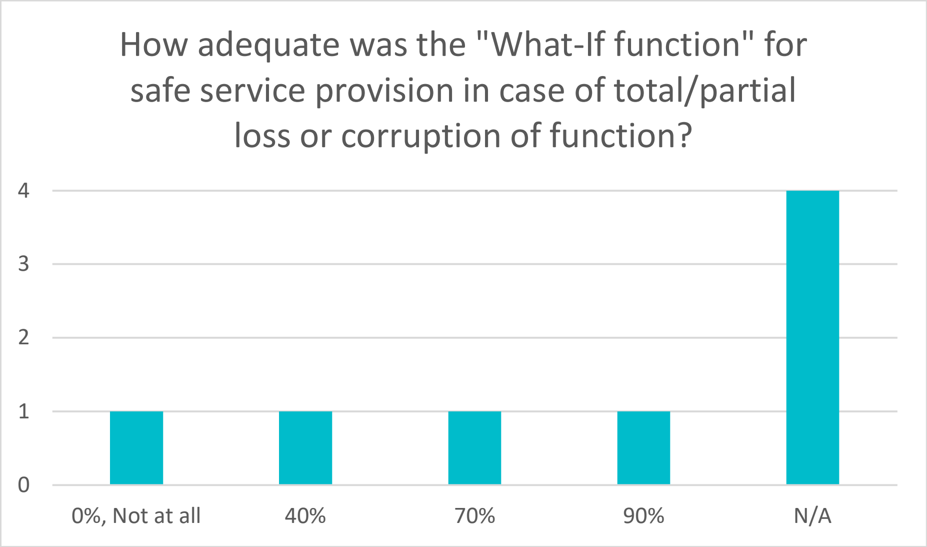


Figure 62. Safety of the logic feedback on "'What-if' function" in abnormal situation.

Objective: To assess safety of system functions in abnormal situations.

Success criteria: The working of conflict resolution advisory was appropriate

The unreliable and sometimes broken functioning of the system made it impossible to make a proper assessment (Figure 63).

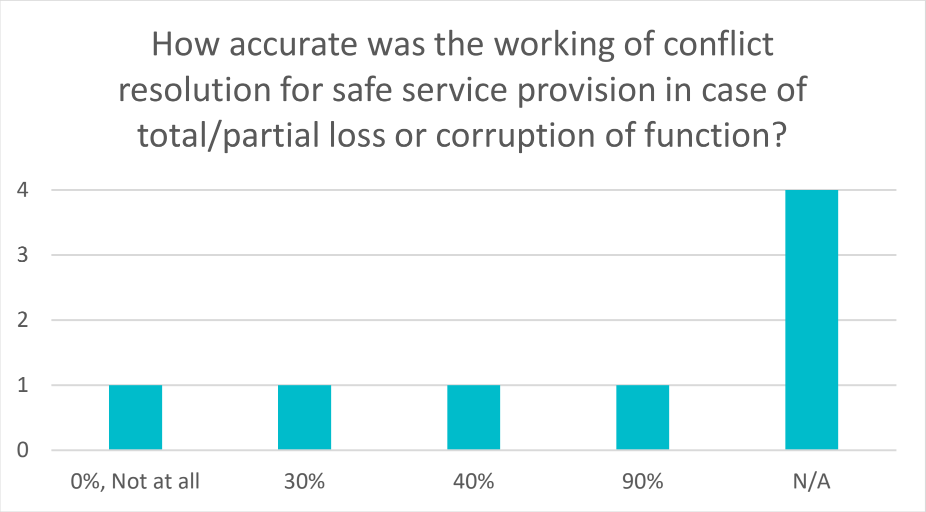


Figure 63. Safety feedback on conflict resolution in abnormal situation.

Objective: To assess safety of degraded modes of system functions.

Convergent feedback from ATCOs indicates that the assessment of the degraded mode objective has failed due to limitations of the system.

4.2.3. HUMAN PERFORMANCE

4.2.3.1 WORKLOAD

The workload was evaluated for real time simulations (RTS) only.

4.2.3.1.1 WORKLOAD IN RTS

Table 28. Table 23. EXE-002- Mental Workload Results.

Objective Criteria ID	Exercise Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
HUM–GREAT–CRT-05	MergeStrip in general reduces the ATCO workload. New functionalities do not increase ATCO workload	The new what-if functionality did not increase the workload. However, the new version did not reduce the workload either.	POK
HUM–GREAT–CRT-05	The what-if function reduces the cognitive workload by supporting the ATCO to find the most optimal solution.	Only 50% of the ATCOs agreed that the what-if function would reduce cognitive workload.	POK

In the first iteration, workload has been addressed after each run and after the whole simulation session. In the second iteration, only post-simulation questionnaire has addressed workload. Figure 64 shows the post-run results between the reference and solution scenarios. The median values are the same in both scenarios (MDN=3).

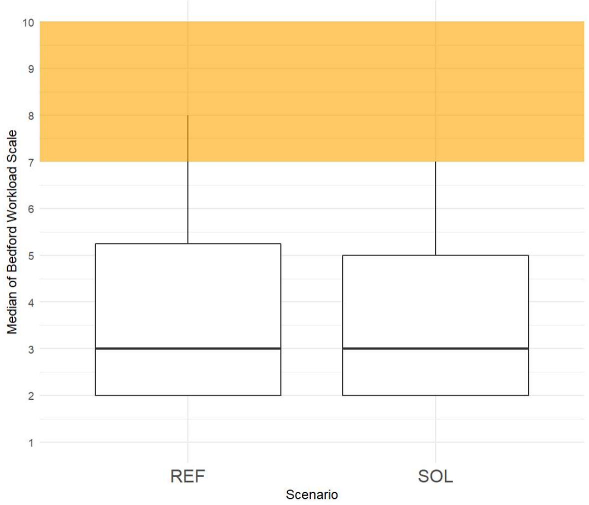


Figure 64. Median values of the Bedford Workload Scale, separated into the reference scenario (current MergeStrip) and the solution scenario (new MergeStrip, what-if function).

The post-simulation questionnaire result is shown in Figure 65. It seems that ATCOs tended to more agree with the statement that the what-if function decreased their cognitive workload after the second iteration, when their previous feedback has been considered. Still, the positive opinions equal with the negative experiences.

ATCOs noted that MergeStrip is only useful when there are only a few arrivals, so the Planner Controller has sufficient time to try different options with the what-if function and analyse the prediction of the sequence order.

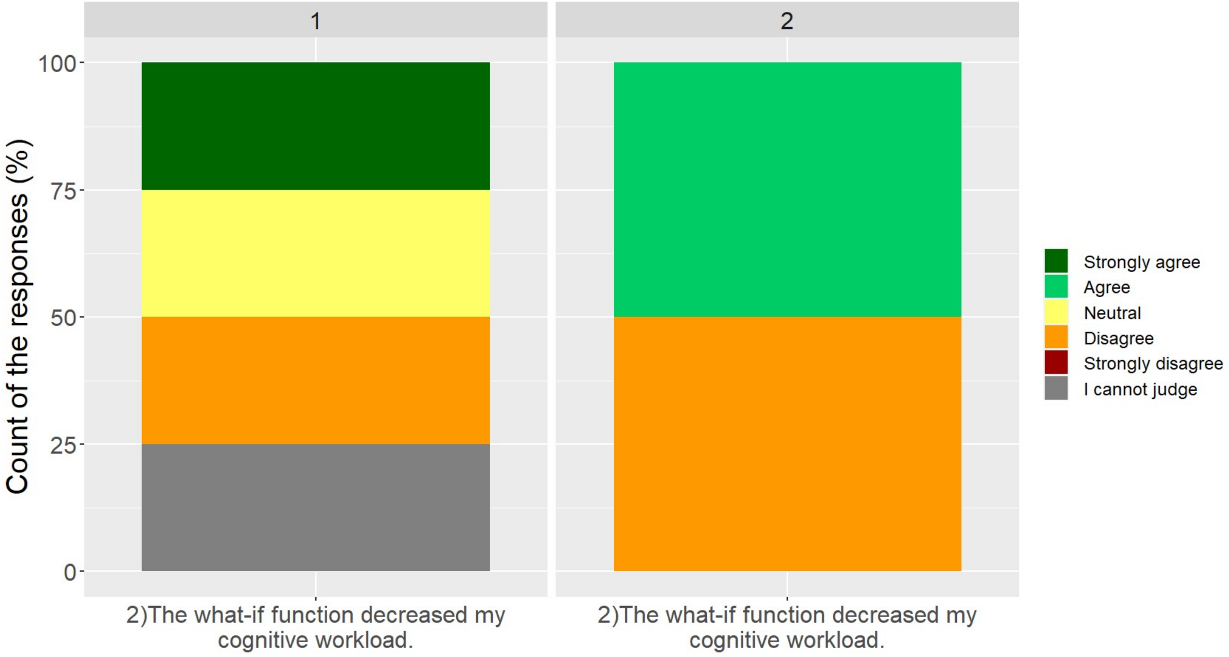


Figure 65. Post-simulation question on cognitive workload in the first and second iteration.

4.2.3.1.2 WORKLOAD IN SHADOW MODE

Workload was not assessed in the shadow mode validation trials.

4.2.3.2 SITUATIONAL AWARENESS

The situational awareness was evaluated for real time simulations (RTS) only.

4.2.3.2.1 SITUATIONAL AWARENESS IN RTS

Table 29. EXE-002- Situational Awareness Results.

Objective Criteria ID	Exercise Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
HUM–GREAT–CRT-06	The what-if function enables ATCO's to make decisions more efficiently.	The majority of ATCOs agreed that the what-if function supports decision-making.	OK

Situational awareness and decision-making have been measured in the post-run questionnaire's SASHA-Q questionnaire (first iteration only), and in the post-simulation questionnaire. SASHA-Q is a standardized questionnaire, and its items address the three different aspects of SA, that is, information extraction, integration and anticipation. SASHA comprises 6 items, which are not assigned to individual scales. Responses to the items are given on a seven-point Likert scale ranging from "never" to "always".

Figure 66 shows the median values of the SASHA-Q average scores, broken down into the reference and solution scenarios. The figure indicates that there is a slight difference in the median values, and the reference solution's value is higher (MDN=5.85, SD=0.98), than the solution scenario's value (MDN=5.5, SD=0.8).

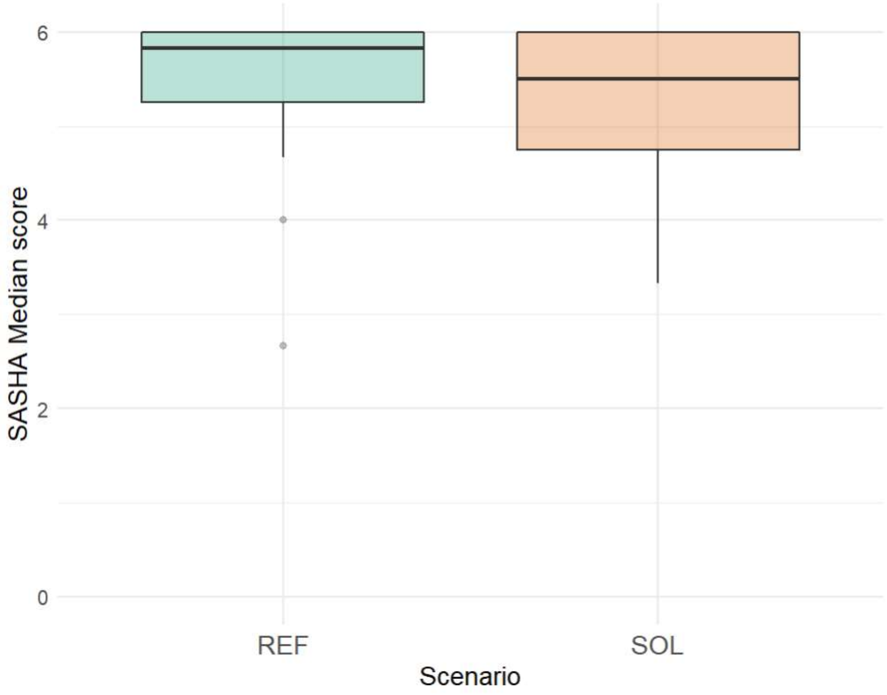


Figure 66. Median values of the SASHA-Q situational awareness score for the reference and the solution scenario.

The post-simulation questionnaire addressed specifically MergeStrip's impact on decision-making. Based on Figure 67, the majority of ATCOs agreed that the new MergeStrip with its what-if function supported the decision-making process.

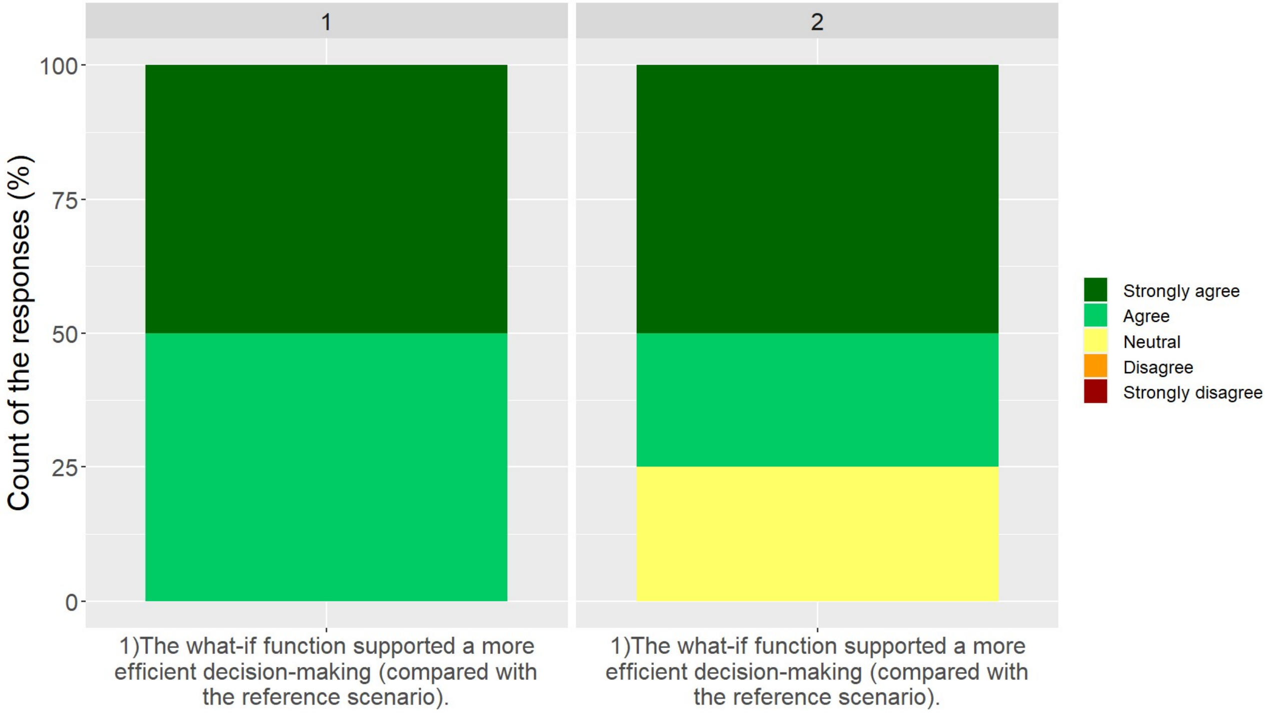


Figure 67. Post-simulation questionnaire about decision-making in the two iterations.

However, it is important to note that above a certain number of arrivals in the TMA, the Planner controller has no capacity to work with the MergeStrip, given that it's on a separate screen in respect to the main ATM system. This is especially the case for the what-if function, where active interaction is needed (e.g. selecting waypoints to probe different sequences). There were no exact number for this capacity limit, but the subjective impression ranges around five. MergeStrip seems to be the most useful in medium traffic density, when the arrivals come from various directions.

4.2.3.2.2 SITUATIONAL AWARENESS IN SHADOW MODE

Situational awareness related objectives were not within the scope of the passive shadow mode.

4.2.3.3 PERFORMANCE OF THE TECHNICAL SYSTEM (I.E., USABILITY, TRUST)

Technical performance was evaluated for real time simulations (RTS) and shadow mode trials.

4.2.3.3.1 PERFORMANCE OF THE TECHNICAL SYSTEM IN RTS

Table 30. EXE-002- Performance of the technical system Results.

Objective Criteria ID	Exercise Validation Success Criteria	Validation Result	VALIDATION OBJECTIVE STATUS
HUM–GREAT–CRT-06	Number and/or severity of errors in the solution is within tolerable limits.	All of the ATCOs agree that the new MergeStrip version did not have a negative impact on human error.	OK

HUM–GREAT–CRT-06	The what-if functionality is easy to interact with.	The interaction got much easier in the second iteration, where the speed probing worked according to the ATCO's mental model. Yet there are still ways to improve the design (e.g. relative speed values, what-if window size)	POK
HUM–GREAT–CRT-06	The look-and-feel of the HMI is acceptable for the ATCOs.	The majority of ATCOs had a positive opinion on the what-if function design by the end of the second iteration.	OK

The usability and user interface design related matters were addressed by questionnaires, observations and debriefing sessions. A great amount of ideas for improvement has been collected, which can be seen in Table 31. The most critical aspects were the what-if window (i.e. size, contents, placement) and the speed probing. Figure 68 shows the design of the what-if window in the first iteration.

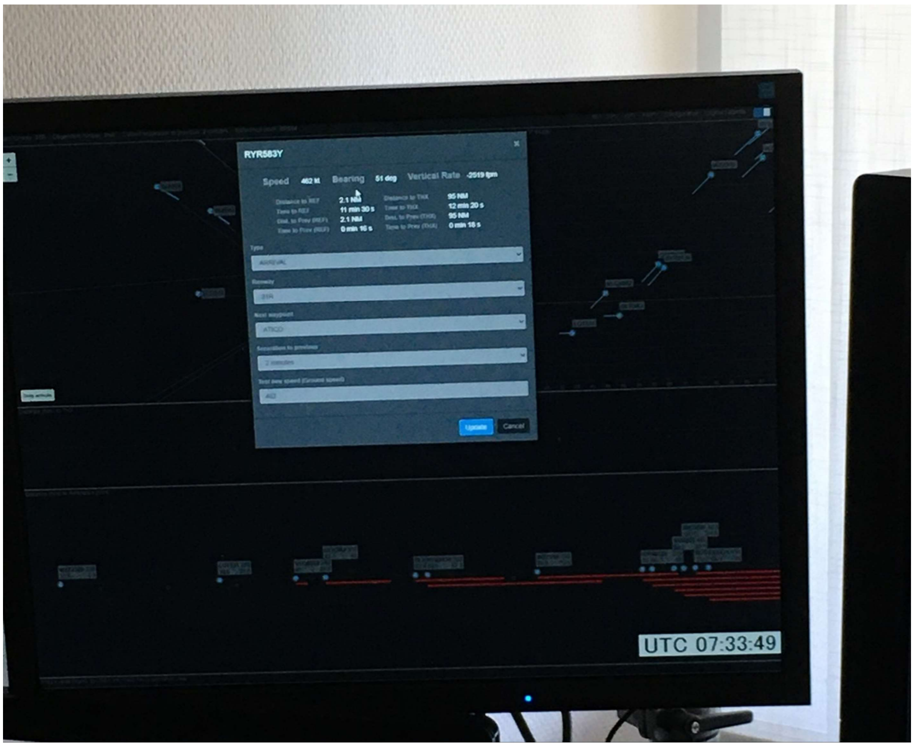


Figure 68. What-if window appearing at the top of the screen, with many unnecessary information on the top.

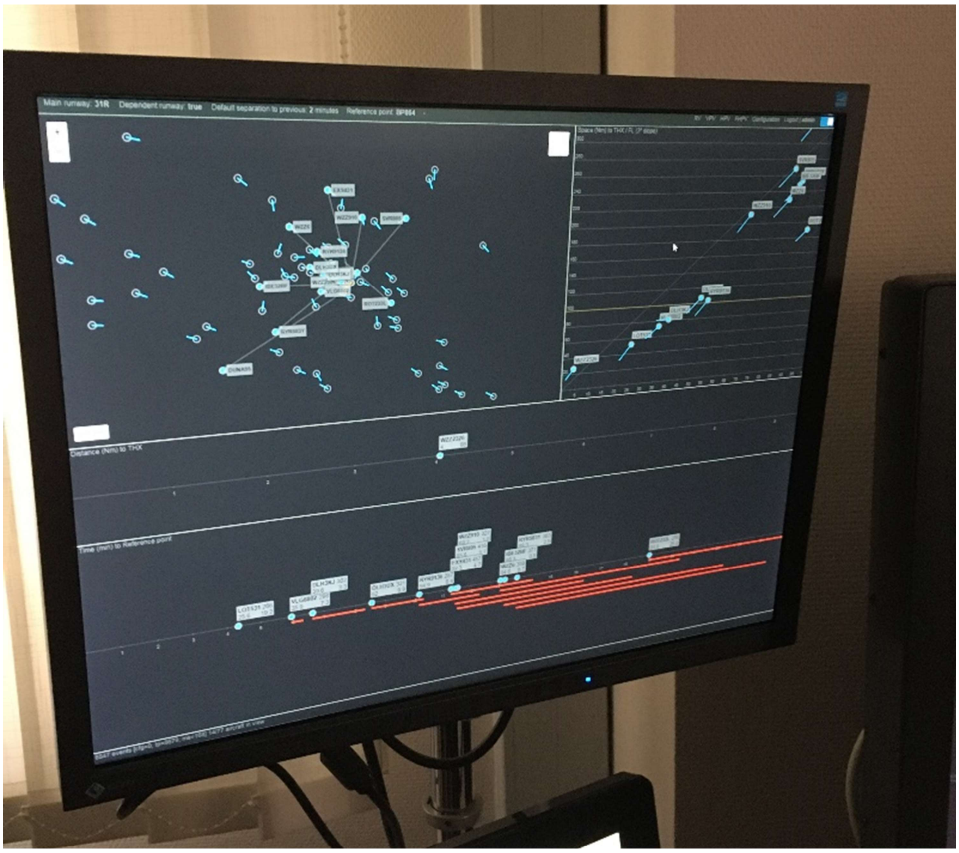

Main runway: 31R
Dependent runway: true
Default separation to previous: 2 minutes
Reference point: BP864
Space (Nm) to THX / FL (3° slope)
Distance (Nm) to THX
Time (min) to Reference point

Figure 69. MergeStrip as a whole in the first iteration, with the red lines that are mentioned in the feedback table.

Table 31. Feedback received from the first iteration.

ID	Main category	Topic	Problem statement	Solution idea
1	what-if: STAR	What-if window: see all the STARs	ATCOs missed some points from the waypoint list. They wanted to check what would happen if the aircraft remained in the STAR, but some points weren't available in the list.	ATCOs would need all the points related to the STARs. To avoid clutter (i.e. too many points in the list), it would make sense to group the points by the STARs and then they can select from only those points that are related to the chosen STAR. Keep only the runway direction, STAR, waypoints and speed dropdown menus in the what-if popup window (in this exact order). Move the "Type" dropdown next to the callsign. Remove the data under the callsign to make the popup window smaller.
2	what-if popup	What-if window size	The window is too large, with unnecessary information (i.e. speed, bearing, vertical rate and the other infos in the upper part).	Redesign the what-if window to make it smaller (see cell E2)
3	what-if popup	Window placement	The ATCO interacts with the system in the bottom, and the popups and infos appear in the upper field (e.g. apply change/cancel, "operation successfully updated"). Thus, they have to drag the mouse cursor away from their main interaction area which also directs the attention elsewhere. Also, there is a small X appearing over the apply change/cancel area which is also disturbing.	The what-if popup window should appear next to the click (the mouse cursor, where their attention is). The info boxes and apply/cancel should also be located in the lower area.

ID	Main category	Topic	Problem statement	Solution idea
4	speed	What if window: Indicated speed	Whilst MergeStrip calculates with the ground speed, ATCOs work with the Indicated airspeed and this is what they communicate to the flight crew. Even if they select a ground speed value that seems to be OK, they don't know how that coverts to IAS and what they should communicate to the pilots to achieve the desired outcome.	It would make more sense to rename & reconfigure the dropdown menu from Ground speed to Indicated airspeed, so that ATCOs can select from IAS. In addition, a small note would appear below showing that this selected IAS would equal to ….ground speed.
5	speed	Dynamic probing	During probe ATCOs had to click to preview change, then cancel and probe again until they found a good solution. It was time-consuming.	It would be more efficient to run probe as ATCOs are selecting the values in the what-if window (this is the desired behaviour compared to row 6). The dot would move as ATCO is scrolling for the optimal indicated speed value. Once it hits the end of the red line, ATCOs would choose that IAS value and communicate it to the flight crew. Thus they would only need to accept the change once and should not click on preview all the time.
6	Other	Screen "jumping"	When an information box appears with green, the screen moves up-and-down (e.g. "operation successfully updated", "preview mode activated"). This is disturbing for the ATCOs.	The appearance of the info boxes should not lead to this jumping effect. These could be moved to the bottom right corner.

ID	Main category	Topic	Problem statement	Solution idea
7	Other	Runway direction change: Selecting the aircraft	Changing the runway direction did not automatically change the route, thus ATCOs had to change the RWY direction for each aircraft individually, which was time-consuming. Also, it was difficult to visually differentiate which aircraft had the correct RWY direction already, so occasionally there were a few that got left out unintentionally.	ATCOs could select a/c by right clicking on their labels. The flow would be the following: 1) Right click on the labels that should REMAIN with the original runway direction (due to them being a smaller number then the ones that need to shift to the new RWY direction). 2) ATCOs change the global runway direction, which results in a change for only those aircraft that were NOT selected by this right click in step 1.
8	Other	Speed what-if function	It's possible that some ATCOs would not see the sense of probing the speed, only the what-if function.	It should be configurable to include or not include the speed into the what-if.
9	Other	Probe: nothing happens after applying the change	It was confusing for the ATCOs that nothing happened after they probed a speed.	See the row in Dynamic Probing. The yellow line is unnecessary during probing.
10	Other	Preview the route		The system should show the route to the new waypoint in yellow.
11	Other	weather mode		In case of bad weather conditions (e.g. thunderstorm), ATCOs would need all the points in the list.
12	Other	conflict lines	ATCOs were not able to follow/comprehend the red lines.	Displaying the conflict lines should be configurable.
13	Other	conflict lines: final	The conflict lines are not of interest in the Distance to THR bay.	Remove the conflict lines completely from the final bay (Distance to THR).

ID	Main category	Topic	Problem statement	Solution idea
14	Other	TMA map colours	The different colours were disturbing.	The map should be the same as in the current MS (TMA and FIR borders + adjacent FIRs, black and white colours).
15	Other	Default setting of labels		Last row in the label: Distance to previous at REF, Distance to previous at THR.
16	Other	Inactive HMI during preview	The system is inactive during preview mode and they cannot interact with it.	

Table 32. Evaluation results whether the changed function works and whether it is acceptable for the controllers.

Functionality	Feedback ID	✓ / ×	Comments
Data reception (Asterix CAT 62 parser)	-		
HMI: FIR & TMA layers overlapping	14		
Automatic change of next waypoint	-		
Manual change of next waypoint: • Without change of STAR • With change of STAR	-		
STARs: • All STARs in the list	1, 11		

Functionality	Feedback ID	✓ / ×	Comments
• All WPs in the list of each specific STAR			
Show/hide all STAR waypoints	1, 11		
What-if window size, position & content	2, 3, 4		
Speed dynamic probing (preview)	5		
No screen jumping	6		
RWY change keeping some A/C in the old RWY (right-clicking)	7		
Switch between *dependent* and *non-dependent* runways	-		
RV: show route to new next waypoint in the preview (yellow color)	10		
HPV: Show/hide conflict lines	12		
Multiple changes in the preview mode	16		

ATCOs received the same post-simulation questionnaires that they had filled out in the first iteration. Figure 70 illustrates the results for both iterations. It seems that the interaction with the system and the overall look-and feel got better for the second simulation (e.g. speed probing).

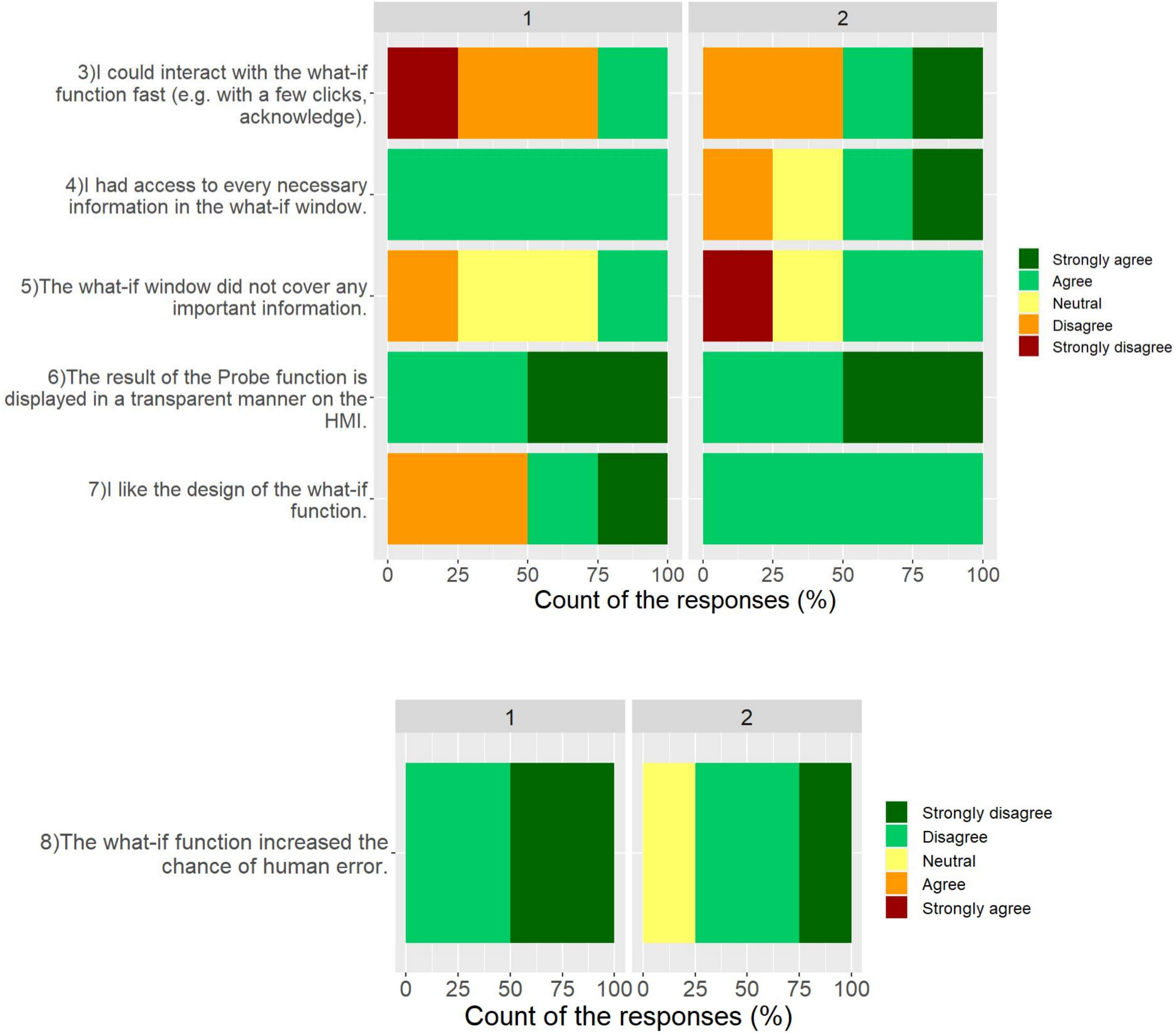


Figure 70. Usability related questions from the post-simulation questionnaire, broken down into the two iterations.

Although their feedback on the what-if function has been integrated into the system, the new window design did not adhere to most of the ATCO's expectations. The same applies to the speed probing: although the interaction with the probe function got easier, the logic behind the speed change remained too complex.

The ATCOs explained their improvement ideas during debriefings, which is summarized in Table 33 created by Pildo Labs, which tracks each remark and categorizes them into different branches.

Table 33. Feedback received from the second iteration.

COLUMN	**Options**			
Feedback origin **- V2D1:** Validation round 2, Day 1 **- V2D2:** Validation round 2, Day 2	*V2D1*	*V2D2*	*V2D1, V2D2*	*Other*
Status	*To be reviewed*	*Consolidated*	*Implemented*	
Priority	*Critical*	*High*	*Medium*	*Low*

ID	Description of the bug / proposed change	Feedback origin	Type	Status	Priority
1	Some flights move backwards in the projection views during the last part of the descent	V2D1	Bug	To be reviewed	Critical
2	WHAT-IF: if the button "Update" is clicked too quick after "next WP" or "speed" is changed, there is no time to load the preview. When this happens, the preview is loaded after the update and the button "apply changes" must be clicked to make it disappear.	V2D1, V2D2	Bug	To be reviewed	Critical
3	RWY change: sometimes works and sometimes not.	V2D1	Bug	To be reviewed	High
4	RWY change: the information in the top menu is not always updated after changing the RWY	V2D2	Bug	To be reviewed	High
5	Temporary overlapping of labels in the HPV	V2D1	Bug	To be reviewed	Low
6	WHAT-IF: When an ATCO accepts or cancels a change of next WP, the yellow line is not removed from the RV of the other ATCO	V2D1	Bug	To be reviewed	Critical
7	WHAT-IF – Bug in preview: change next WP (do not accept neither cancel the change), click the label of the previewed projection, click update (the original flight disappears), click "Cancel" in the preview list (the previewed flight also disappears)!!! Solution: preview label not clickable + change background color of modified parameter(s) + if operation window is closed (X) keep the non-applied changes when it is reopened	Other	Bug	To be reviewed	Critical

8	The size of the operation window should be even smaller (reduce spaces and/or font size)	V2D1, V2D2	Change	To be reviewed	
9	WHAT-IF: Dynamic speed probing: use relative speeds (changes of +-10kts) instead of absolute values of GS/IAS	V2D1	Change	To be reviewed	
10	WHAT-IF: Remove buttons "Apply changes"/"Cancel" from the preview changes list. Changes shall be only accepted/discarded from the operation window	V2D1	Change	To be reviewed	
11	TMA overlay: show only the external borders of the TMA (remove the divisions within the TMA)	V2D1	Change	To be reviewed	
12	2 waypoints are missing ("behind" the Tbar). They are not part of any STAR. They should appear on the list of Wps independently on which is the selected STAR	V2D1	Change	To be reviewed	
13	Missing Waypoints: BP744, BP774	V2D1	Change	To be reviewed	
14	WHAT-IF: Yellow line in RV (preview) should be extended until the RWY. Make it a little thicker	V2D1, V2D2	Change	To be reviewed	
15	Labels in the RV should be draggable	V2D1	Change	To be reviewed	
16	The reference points should be the merge points (IF, Tbar central WP)	V2D2	Change	To be reviewed	
17	Operations close to the THX should not disappear	V2D2	Change	To be reviewed	
18	RV: lines from AC to WP should be always displayed independently on the configured RWY	V2D2	Change	To be reviewed	
19	New way to detect change of next WP: if the system detects that a specific AC moves away from its next WP, trigger the process to detect a new next WP. EXAMPLE. Consider the following WP sequencing: [A, B, C, D, E, REF_POINT]. Consider that the current next WP is B. When the AC moves away from B, use next algorithm to select next WP: IF A/C approaches C → next WP = C ELSEIF A/C approaches D → next WP = D ELSEIF A/C approaches E → next WP = E ELSE → next WP = REF_POINT	V2D2	Change	To be reviewed	

20	Use persistent database to store Users and user configuration settings. Also RWYs, STARs & WPs	V2D2	Change	To be reviewed	
21	Apply a visual distinction to flights following the STAR	V2D2	Change	To be reviewed	
22	HPV: currently, in order to see all flights in the HPV, a large zoom out must be applied. In this case, most of the flights are accumulated in a short portion of the strip, which is not optimum. Enable the user to scroll horizontally along the strip to see all flights without requiring to apply the minimum zoom level.	V2D2	Change	To be reviewed	
23	Handle missed approaches: do not remove a flight when it arrives at the THX (and is considered to be finished). Instead, hide it from all views and, in case we still receive data from it after X seconds and its height is not 0, consider it as a missed approach and show it again.	V2D2	Change	To be reviewed	
24	Enable the user to define a custom route to the REF_POINT. The system should allow to remove some of the points of the STAR for a specific flight, thus making the route shorter.	V2D2	Change	To be reviewed	

4.2.3.3.2 PERFORMANCE OF THE TECHNICAL SYSTEM IN SHADOW MODE

Objective	Success criteria	Result	Status
HUMAN PERFORMANCE – USABILITY			
To assess the usability of the system.	The improved ETA prediction supports more efficient task performance (arrival sequencing).	Judging the ETA improvement was heavily impacted by the instability of the system.	Not applicable, could not be properly tested.
	The conflict resolution advisory supports efficient task performance by avoiding non-optimal tactical intervention (i.e., vectoring, holding)	The impression was that the Recommender function was inconsistent and the advisories were often unjustified.	NOK
	Number and/or severity of errors in the solution is within tolerable limits.	ATCOs' feedback indicates that neither new functions increase the chance of human error,	OK
	The 'What-if' functionality is easy to interact with.	The window was further improved by reducing the content to the most crucial information. It was easy to visually distinguish the probed and original aircraft, and the test new speed also worked as expected.	OK
	The conflict resolution advisory function is easy to interact with.	ATCOs confirmed in the debriefing that it was easy to visually distinguish the original label, the probed aircraft and the aircraft which had recommendations. The feedback however is affected by the functionality showing unjustified suggestions.	POK
	The look-and-feel of the HMI is acceptable for the ATCOs.	The colours, readability and general interaction with the system was acceptable.	OK
HUMAN PERFORMANCE – TRUST			

To assess the ATCO's trust in the system.	ATCOs trust in the accuracy of the new ETA prediction.	Judging the ETA improvement was heavily impacted by the instability of the system.	Not applicable, could not be properly tested.
	The conflict resolution advisory provided by the system is perceived sensible by the ATCOs. // The conflict resolution advisory provided by the system fits the ATCO's expectations.	The Recommender functionality did not work as expected and often suggested unrealistic options. Therefore, only around 20% (2 out of 9) of the participants agreed with the statement in the success criteria.	NOK

Objective: To assess the usability of the system.

Success criteria: The improved ETA prediction supports more efficient task performance (arrival sequencing).

In general, ATCOs were hesitant to draw any conclusions on the accuracy of the ETA calculation, since the system often froze after ~30 minutes. Anyhow, it seemed that it did consider the future aircraft speed (i.e. the arrival will reduce speed), so it became obvious that the algorithm is not just based on the current speed, which was welcomed (Figure 71). Over the shoulder observations showed that improved ETA had no real added value in more efficient task performance in a way that this functionality was already known by them.

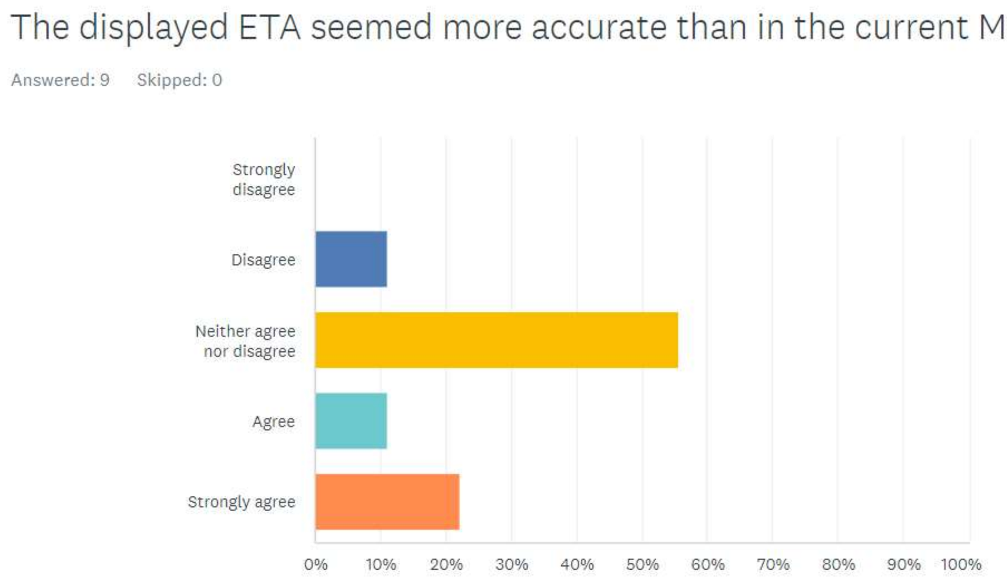


Figure 71. HF feedback on ETA accuracy.

Success criteria: The conflict resolution advisory supports efficient task performance by avoiding non-optimal tactical intervention (i.e., vectoring, holding)

According to the feedback the Recommender function was not sufficiently mature to support ATCO performance. Most of the time it suggested to send the aircraft to BP854, which the ATCOs cannot do. There were some occasions when two aircraft were close and the system aimed to create more space between them by sending the second one to another waypoint. This was deemed as an interesting attempt, however, after 1 minute the system changed its mind and suggested the opposite. (Note that this situation would have been solved by the ATCOs by adjusting the speed). In general, it seemed as if it did not consider the speed of the aircraft. For instance, after picking one aircraft to slow down, it did some recalculation and realised that it was faster originally so this was not a good solution, and then suggested the opposite. The impression was that the Recommender function was thus inconsistent and decreased the potential to build trust in the system (Figure 72).

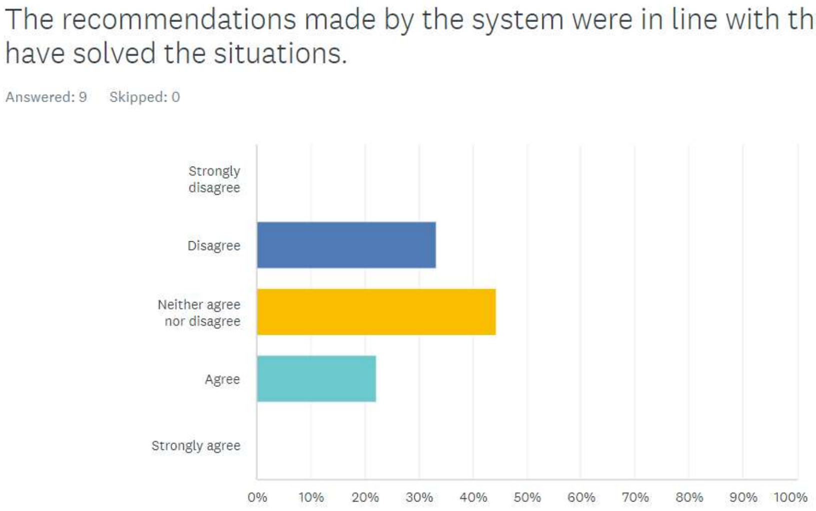


Figure 72. HF feedback on the recommender function.

Success criteria: Number and/or severity of errors in the solution is within tolerable limits.

ATCOs' feedback indicates that neither new functions increase the chance of human error, which sends a positive message (Figure 73).

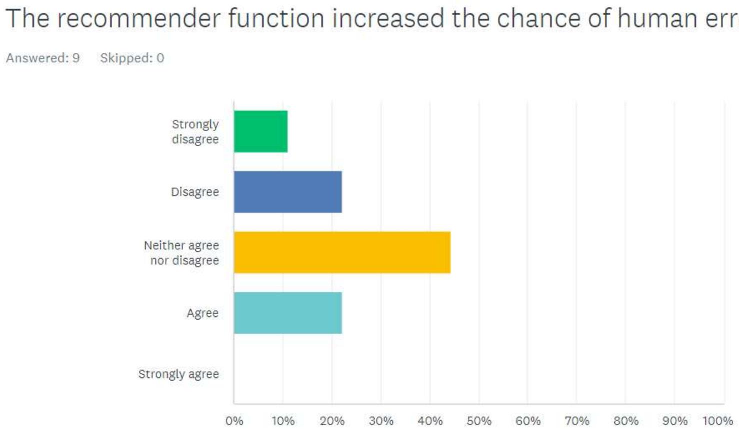


Figure 73. HF feedback on human error.

Success criteria: The 'What-if' functionality is easy to interact with.

First, and not necessarily related to only the 'What-if' function, ATCOs were asked in general to judge the interaction with the system. Based on Figure 74, the majority of the participants rated the interaction as efficient.

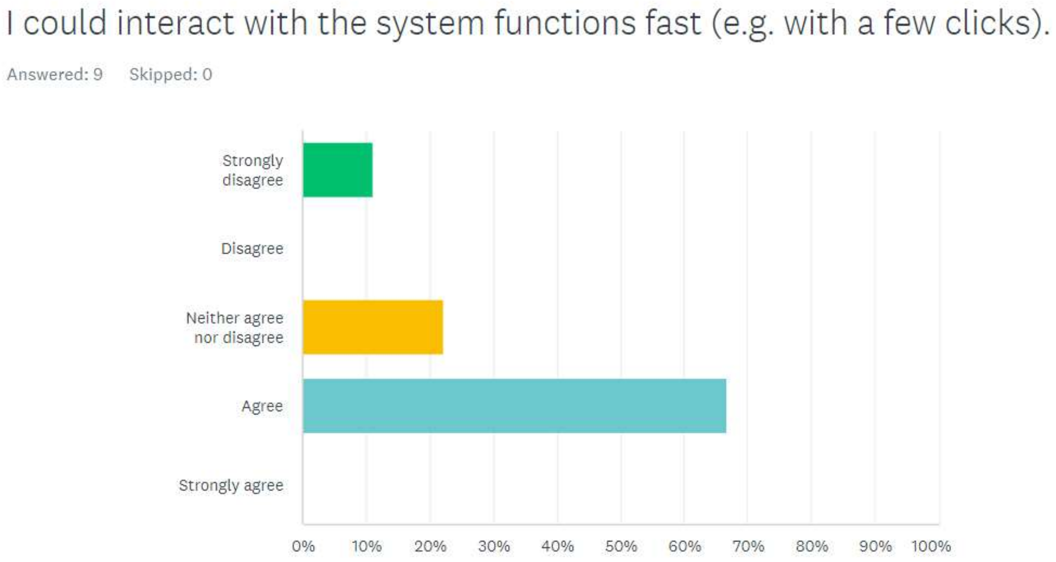


Figure 74. HF feedback on the interaction with the system.

In terms of the 'What-if' function, its window was further improved by reducing the content to the most crucial information (i.e. Type, STAR, Waypoint, Speed) and this is attested by the feedback (Figure 75). Furthermore, according to the opinions the probed aircraft was well distinguished from the original track and label (Figure 76). Lastly, the look-and feel of the design was also touched upon, which the majority of the ATCOs liked (Figure 77). The font sizes and colours well acceptable, although the contrast between white letters and yellow background was too low to ensure readability. Over the shoulder observation can confirm that these answers reflect the difference of testers' perception between the beginning and the end of the validation trial.

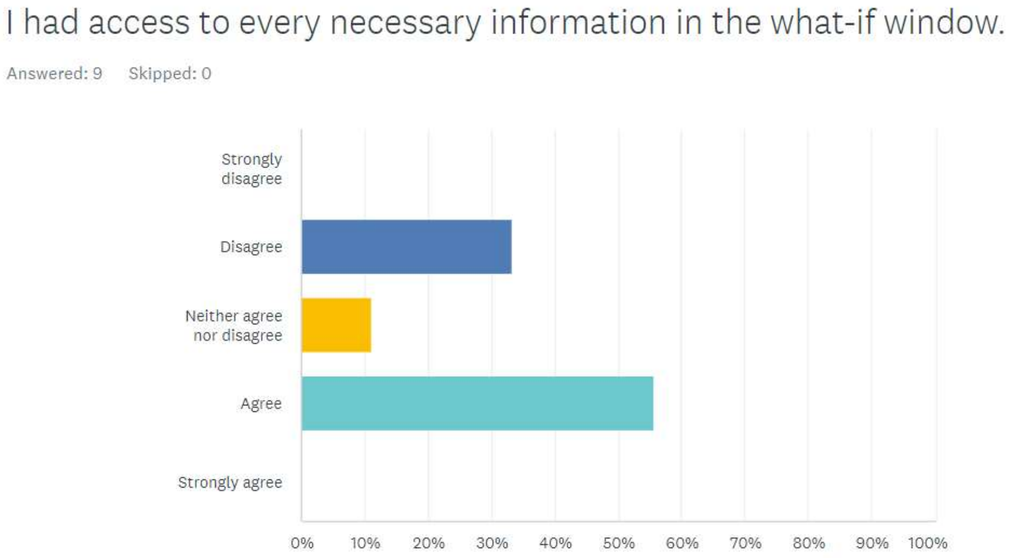


Figure 75. HF feedback on the 'What-if' window content.

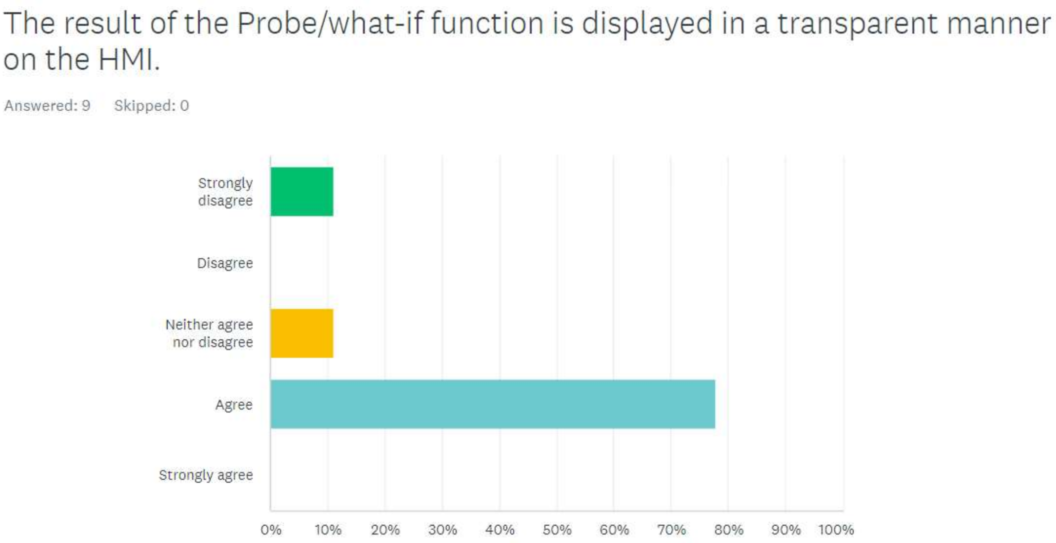


Figure 76. HF feedback on the 'What-if' transparency.

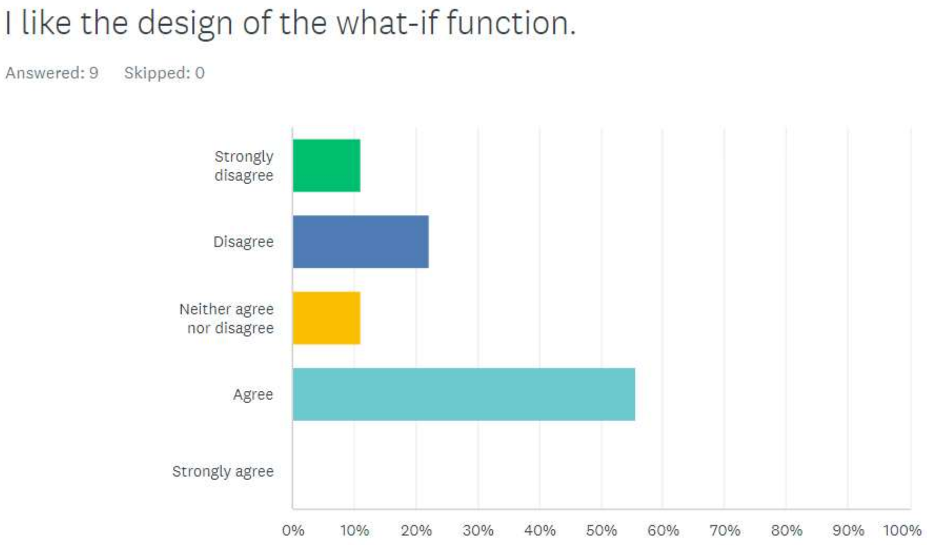


Figure 77. HF feedback on the 'What-if' design.

Success criteria: The conflict resolution advisory function is easy to interact with.

As Figure 78 suggests, around half of the participants agreed that the recommender design is agreeable. Although the functionality did not work as expected, ATCOs confirmed in the debriefing that it was easy to visually distinguish the original label, the probed aircraft and the aircraft which had recommendations (i.e. green label).

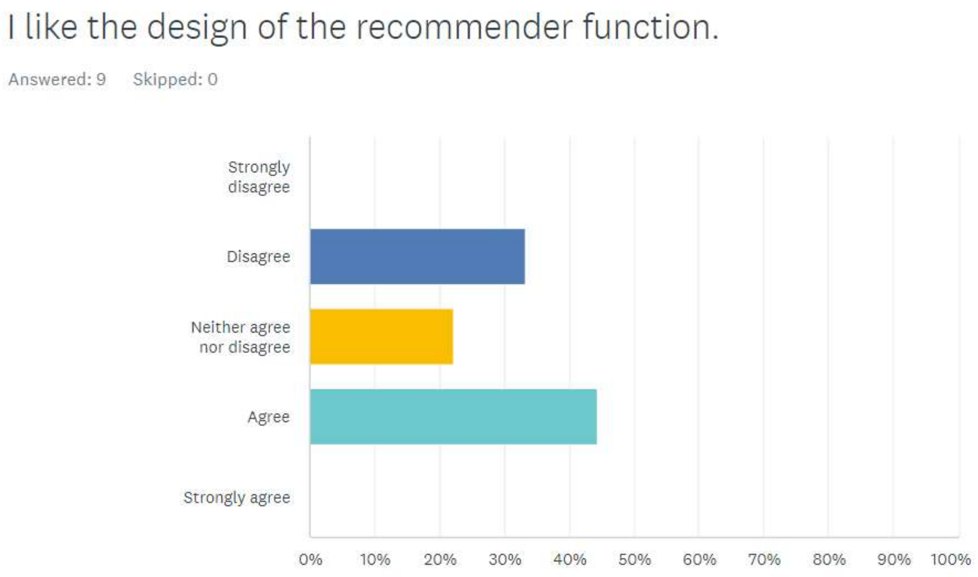


Figure 78. HF feedback on the recommender function.

Success criteria: The look-and-feel of the HMI is acceptable for the ATCOs.

The debriefing session touched upon the aesthetics of the HMI. ATCOs confirmed that the colours, font size and the general interaction with the system is good. The only thing that came up that the 'What-if' function marks the label of the probed aircraft with a yellow background, and in this case its contrast with the white font leads to reduced legibility.

Objective: To assess the ATCO's trust in the system.

Success criteria: ATCOs trust in the accuracy of the new ETA prediction.

This success criteria is difficult to address, as during the validation the system often froze. Even though the ATCO choose an arrival to monitor its ETA, MergeStrip often froze and comparison was impossible. Therefore, ATCOs unanimously suggested to improve the system's stability, otherwise trust cannot be built.

Success criteria: The conflict resolution advisory provided by the system is perceived sensible by the ATCOs. The conflict resolution advisory provided by the system fits the ATCO's expectations.

As described earlier, the Recommender functionality did not work as expected and often suggested unrealistic options. Therefore, only around 20% (2 out of 9) of the participants agreed with the statement posed in the questionnaire (Figure 79).

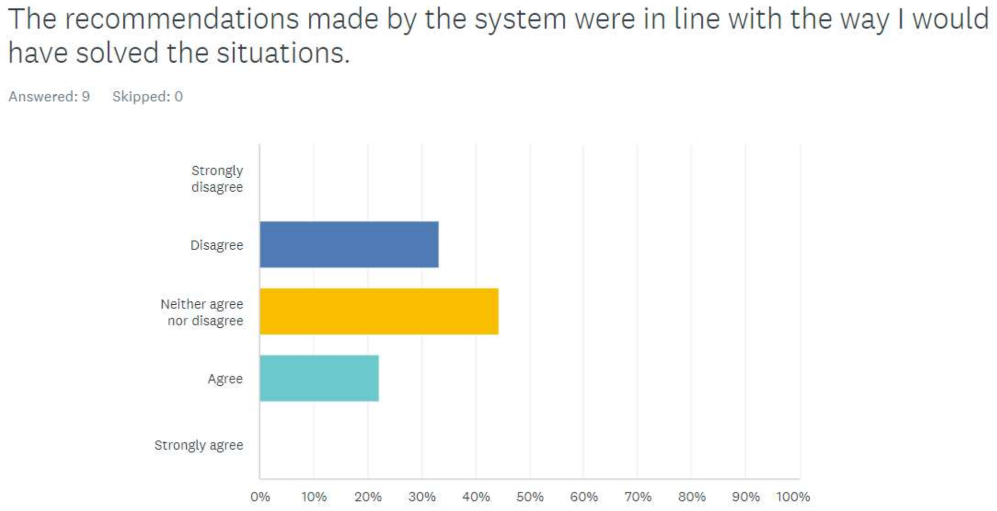


Figure 79. HF feedback on the recommender function (same as Figure 72).

As a general note on trust, it is worth to analyse the matter by listing some of the components of trust.

ATCOs had the possibility to fill out a paper-based questionnaire after working with MergeStrip. Based on the notes it became obvious that one of the key cornerstones, i.e. **reliability** was heavily impacted by the instability of the system.

However, in terms of **usefulness**, we can summarize that:

- The 'What-if' function's „Test New Speed" was working well and it was useful to immediately see the change.
- The system showed the one-minute speed vectors on the glide slope; thus the rate of the descent can be seen at a glance.
- In high traffic load it helps a lot that the arrivals appear automatically.
- It looks farther than the current MergeStrip.
- It's great that it calculates with all the waypoints of the STARs.

The main issue was that a basic functionality (i.e. Threshold separation tool) was unusable as the system had the error of turning back the arrivals at 5NM prior to landing. This malfunction was frustrating for the ATCOs.

With regards to the **accuracy**, it seemed that the "distance to previous" value was more accurate, and it was welcomed that the system did not only calculate with the current speed.

When it came to **understandability**/transparency, as mentioned before, the recommender's suggestions simply did not work and were often seen as unjustified.

4.2.4. CAPACITY

4.2.4.1 NUMBER OF ARRIVALS PER UNIT OF TIME

The number of arrivals were separately evaluated for real time simulations (RTS) and shadow mode trials.

4.2.4.1.1 NUMBER OF ARRIVALS PER UNIT OF TIME IN RTS

Each simulation exercise lasted for 45 minutes. In the first iteration the following number of arrivals were part of the scenarios.

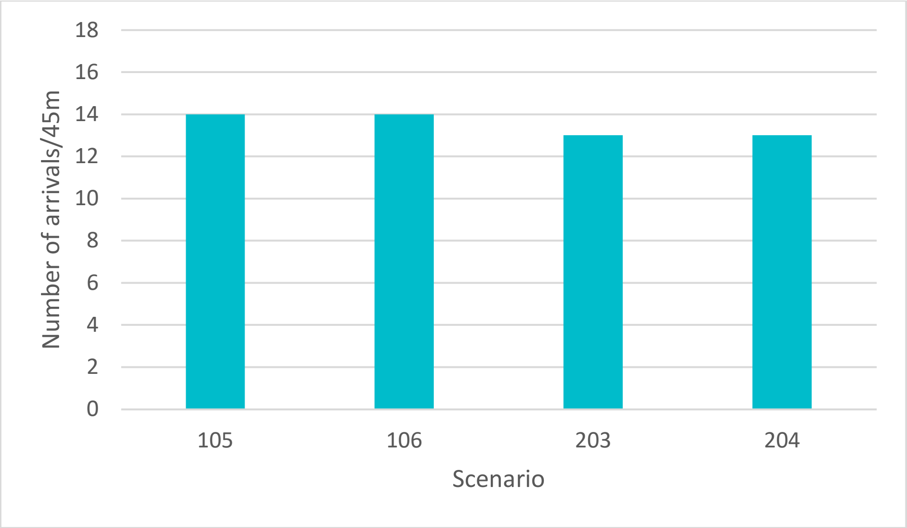


Figure 80. Number of arrivals in the first iteration.

In the second iteration the team decided to apply more scenarios to make sure the ATCOs can test the system with various traffic situations. There were scenarios with less arrivals and others with more arrivals than in the first iterations.

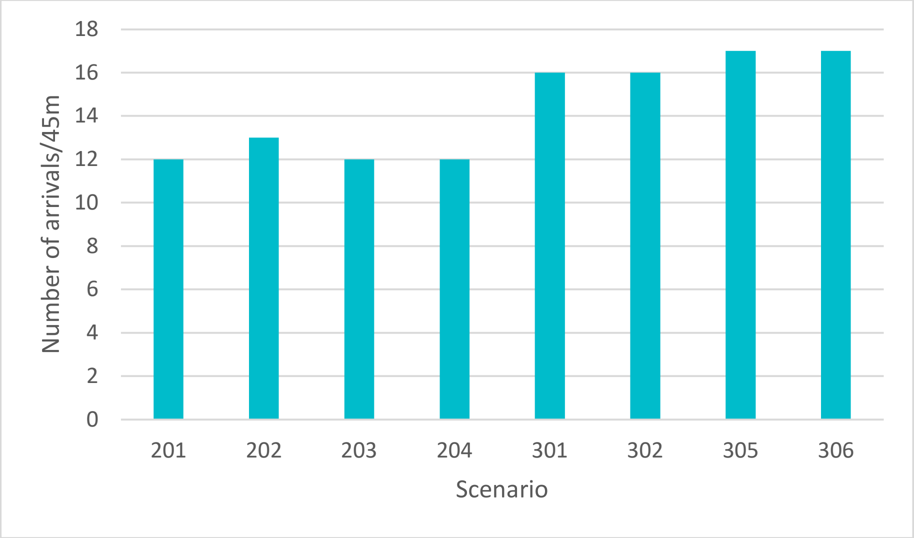


Figure 81. Number of arrivals in the second iteration.

To summarize the ATCOs' feedback, the MergeStrip with its what-if function is not there to increase capacity. To put it differently, the planner controller needs the capacity to actively interact with the what-if function and probe waypoints and speed values, which cannot happen when the number of arrivals is high in the TMA.

4.2.4.1.2 NUMBER OF ARRIVALS PER UNIT OF TIME IN SHADOW MODE

The trajectories used for this analysis were generated from real ADS-B data recorded in Budapest Ferenc Liszt International airport during the period in which MergeStrip was tested in the OPS room (between March 31st and April 13th). Within this period, MergeStrip was tested during three specific time slots: 09:45-11:30, 15:45-17:00 and 20:30-22:00 (UTC).

As it will be further explained in Section 4.3.3, Budapest TMA traffic is still below the pre-COVID level, and also below what was expected for 2023. During the execution of the shadow mode exercise, the traffic level was far below the maximum capacity of the airport in terms of number of arrivals per unit of time. For this reason, no relevant capacity-related conclusions could be extracted from this validation exercise.

However, over-the-shoulder observation and final user workshop can confirm that the new version enabled handling 7-8 aircrafts at the same time which is a major step compared to the currently used version with 5-7 aircraft.

4.2.5. SUMMARY OF EXERCISES RESULTS

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	VALIDATION OBJECTIVE STAT	
ENVIRONMENT					
ENV – GREAT – OBJ-01	To assess the reduction of exhaust emissions due to solution	ENV – GREAT – CRT-01-10	The solution results in reduction of exhaust emissions compared to the reference scenario.	OK	
OPERATIONAL EFFICIENCY					
OPE – GREAT - 02	To assess the reduction in flown distance per aircraft due to solution	OPE – GREAT – CRT-02-10	The distance flown is reduced compared to reference scenario.	OK	
OPE – GREAT - 03	To assess reduction in fuel burn due to solution	OPE – GREAT – CRT-03-10	The average fuel burn by aircraft is reduced compared to the reference scenario.	OK	
CAPACITY					
CAP – GREAT - 04	To assess the solution's impact on capacity	CAP – GREAT – CRT-04-10	The solution does not reduce capacity.	OK	Increase was experienced in Shadow mode.
HUMAN PERFORMANCE – WORKLOAD					
HUM – GREAT - 05	To assess the ATCO's workload	HUM – GREAT – CRT-05-10	The level of workload is within acceptable limits.	POK	The workload is within acceptable limits, but the exercise validation criteria are only partially met. Was not

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	VALIDATION OBJECTIVE STAT	
					assessed in Shadow mode.
HUMAN PERFORMANCE – SITUATIONAL AWARENESS					
HUM – GREAT - 06	To assess the ATCO's situational awareness	HUM – GREAT – CRT-06-10	The level of situational awareness is within acceptable limits.	OK	Was not assessed in Shadow mode.
HUMAN PERFORMANCE – USABILITY					
HUM – GREAT – 07	To assess the usability of the system	HUM-GREAT-CRT-07-10	There is no discrepancy between system-provided information and user-required information.	NOK	Remaining needs are detailed in Table 33
		HUM-GREAT-CRT-07-20	The ATCO can perform interaction without noticeable problems.	OK	
		HUM – GREAT – CRT-05-40	The look-and-feel of the HMI is acceptable.	OK	
HUMAN PERFORMANCE – TRUST					
HUM – GREAT – 08	To assess the ATCO's trust in the system	HUM – GREAT – CRT-08-10	The level of trust is experienced as sufficient by the ATCO.	NOK	The system was not stable enough to determine reliability, and the accuracy of the ETA calculation. The logic behind the recommendations was difficult to understand and seemed inconsistent.

Objective ID	Validation Objective	Criteria ID	Validation Success Criteria	VALIDATION OBJECTIVE STAT	
SAFETY					
SAF – GREAT – 09	To assess the impact on the safety level of the system	SAF – GREAT – CRT-09-10	Procedures and system functions are safe in normal situations.	NOK	The system was not stable enough to make a proper assessment.
		SAF – GREAT – CRT-09-20	Procedures and system functions are safe in abnormal situations.	POK	Abnormal scenarios were not specifically addressed, but unexpected situation was part of one of the scenario. Was not assessed in Shadow mode.
		SAF – GREAT – CRT-09-30	Procedures and system functions are safe in degraded mode situations.	N/A	Was not assessed in Shadow mode.

4.3. ANALYSIS OF EXERCISE RESULTS

4.3.1. HUMAN PERFORMANCE RESULTS

- **EXE-001 – DLR**

Overall, mental **workload** ranged between medium to low workload. Based on ISA measurement, two conclusions could be drawn. Firstly, mental workload remained at acceptable levels pointing out no mental overload. This finding was also confirmed through controllers' feedback during the debriefing session. Nevertheless, ISA ratings the 60% 4D-FMS simulation run of the first validation iteration and the 80% 4D-FMS simulation run of the second validation iteration pointed towards mental underload. A particular attention should be given to this point as a mental underload could be a potential safety risk. Secondly, the experienced mental workload seemed to be inversely related to the percentage of 4D-FMS aircraft. Increasing the amount of untouchable 4D-FMS aircraft results in a reduction of the share of 3D-FMS aircraft navigated by the ATCO. Indeed, the number of aircraft a controller manages simultaneously at a given time has been the most used index to estimate the workload in many studies, which seems to be in line with the simulation results [Athènes 2002]. However, this index is biased by the way aircraft are spread over space and time [Athènes 2002] and therefore, less aircraft to be managed do not necessarily result in less workload. An alternative explanation could be linked to the main task of the ATCO: Given the route structure (separated by design) as well as the sequence proposed by the AMAN (considering required separation), the ATCO mainly monitored and guided the 3D-FMS aircrafts towards the target window to meet the optimal position on final, unless he/she decided to choose an alternative path based on direct routing. That being said, the higher the amount of untouchable aircraft, the less intervention is required from the controller, potentially resulting in lower mental workload. Putting it all together, mental workload during the trials was interpreted as acceptable. However, it could be expected to be lower in simulations than in real operations. It is then strongly recommended to be tested in real operations using a bigger sample size to ensure that it remains within acceptable limits.

Perceived **situation awareness** remained at an acceptable level for a 4D-FMS aircraft percentage of up to 60%. The ghosts and target windows were seen as beneficial regarding situation awareness, as reported during the debriefing. More research will be needed to assess the impact of higher percentages of untouchable 4D-FMS aircraft on situation awareness. Generally speaking, the degree of operator involvement in a task directly influences situation awareness. Monitoring automated systems and assuming a more passive role instead of actively engaging with a system can impair situation awareness, possibly resulting in an out-of-the-loop performance problem [Endsley 1995]. Because a higher share of 4D-FMS aircraft leads to the ATCO passively monitoring more aircraft, a too high amount of 4D-FMS aircraft might result in lowered situation awareness. This possibility should be critically considered in future research.

The SUS indicated "good" or close to "good" **usability** [Brooke 1996]. It should be also noted that ATCOs had mentioned some challenges working with unfamiliar control working position. They explained that the radar situation display used at home disposes of very large advanced features which they missed during the trials. Thus, the overall usability rating may be impacted by such difference in the CWP. In addition, the ATCOs proposed various improvements to the system (CWP and tools) in regards to the system-provided information, interaction with the system and the look-and-feel of the HMI (Chapter 7.1). This shows that there is room for improvement regarding the usability of the system. The ATCOs' propositions and concerns should be considered for the future development of the system.

The level of overall **trust** into the system was slightly above medium. This was interpreted as a sufficient level of trust. The robustness of the system was identified as an area of improvement in particular. One possible explanation for this could be the fact that several technical issues occurred during the simulations in both validation iterations. For future evaluations of the system, it needs to be ensured that these technical issues will be remedied. The understandability of the system seemed to be good, as this SATI-subscale was rated especially high.

- **EXE-002 – HC**

Real-time Simulation

One of the key criteria the team focused on in the first validation session was **usability and user interface design**. This was to ensure that the system the participants were testing are in accordance with their expectations and supports efficient traffic management. The main points the ATCOs emphasized were the what-if function’s window size, which was too big hence covered important areas. The other functionality was the speed probing, which has been improved for the second iteration massively. Other design elements have also been pointed out and have been taken care of by Pildo Labs for the second simulation (e.g. red lines, runway direction change). The two session so close in time (i.e. September, November 2022) enabled the team to experience the benefit of this agile software development approach and the results illustrates the improving impressions of the system usability.

However, an important factor when using MergeStrip is that it is only optimal in medium traffic density, when the arrivals come from different directions. This is not the outcome of this validation as the ATCOs are already using MergeStrip in its current version. Still, the what-if function adds a new dimension where they have to actively interact with the system and probe different waypoints or speed values to see how that effects the arrival sequence. The planner controller reached its capacity in the high traffic density scenarios and could divide his attention between the main ATM system and MergeStrip. The **workload and situational awareness** results reflect the above-mentioned experience – the majority of ATCOs agreed that the what-if function supported decision-making, but it depends on the actual traffic density.

Shadow Mode Validation

Compared to the previous validation activity, the current one focused on passive shadow mode, which means that an additional ATCO was observing the situation from another CWP in the OPS room. This enabled to see the accuracy and usability of the system when it receives real data feed. In addition, new functionalities have also been added to the system, namely the improved ETA calculation and a recommender function.

As mentioned, the main area of focus was accuracy, usability and user interface design. This was to ensure that the system the participants were testing were in accordance with their expectations and supports efficient traffic management. Unfortunately, the system was relatively unstable to properly test the ETA calculation. Yet it seemed that the algorithm was more advanced than the one in the current MergeStrip and did not take only into account the current speed, but also other characteristics (e.g. arrivals will slow down). The ‘What-if’ function usefulness and usability was further confirmed. However, the new recommender functionality did not live up to the expectations. The suggestions seemed unjustified and inconsistent most of the time.

4.3.2. FEASIBILITY AND ACCEPTABILITY FROM CONTROLLERS’ PERSPECTIVES

- **EXE-001 – DLR**

The concept was generally accepted and perceived as feasible by the ATCOs. Several ATCOs reported that they would like to see the ghosts as well as the target windows implemented in real operations. They clarified that such features could be useful even for other purposes and other concepts

However, the general idea of untouchable aircraft (not under the "control" of the ATCO) was not accepted by all ATCOs. Critical comments regarding the concept of untouchable 4D-equipped aircraft were that this was judged to impair the effectiveness of conflict resolution and discriminate against non-equipped aircraft. Some ATCOs expressed that they felt that the overall concept put 3D-equipped aircraft at a disadvantage. The target windows were also criticized regarding efficiency because they did not add to finding the most efficient solution in the ATCOs' opinions. This might be explained by the fact that the AMAN computed the optimum sequence environment friendly wise and not capacity wise.

- **EXE-002 – HC**

Real-time Simulation

The what-if function of the new MergeStrip was generally accepted and perceived as feasible by the ATCOs, but only in medium traffic density, when the arrivals come from different directions. Some ATCOs also explained that instead of focusing on the arrivals that are close to the final, the system could rather focus the arrivals that are further and could rather help setting up the sequence more ahead in time.

Even after the second iteration the ATCOs gave a lot of feedback on further improving the system. This shows the need to apply more agile system development projects, where ATCOs can test the new version in a realistic situation.

Shadow Mode Simulation

As for the passive shadow validation in the OPS room, the main areas of focus were accuracy, usability and user interface design. Unfortunately, the system was relatively unstable to properly test the ETA calculation. Yet it seemed that the algorithm was more advanced than the one in the current Merge Strip and did not only consider the current speed, but also other characteristics (e.g. arrivals will slow down). The what-if function usefulness and usability were further confirmed. However, the new recommender functionality did not live up to the expectations. The suggestions seemed unjustified and inconsistent most of the time.

4.3.3. ENVIRONMENTAL SUSTAINABILITY & FUEL EFFICIENCY RESULTS ASSESSMENT

The focus of the automated simulation runs for the surface traffic was to improve the optimization of component of SMAN software to generate more environmentally friendly taxi trajectories. To achieve this, the main focus was to eliminate unnecessary holding times. By using a conflict resolution algorithm with a parameterized optimization function, it was possible to use a green profile for the trajectory generation. This profile was configured to prefer holding aircraft at their stand before engine start-up and includes higher penalties for holds during normal taxi. This resulted in a reduction of the number of holds by 80% in high-density traffic scenarios, compared to a conventional optimized trajectory profile. Furthermore, the green profile consistently produced trajectories with fewer holds than the conventional profile, regardless of traffic density and planning times.

Unfortunately, the results of the trajectory analysis could not be used to conduct a meaningful quantification of the estimated fuel consumption. The available model by BADA uses only the taxi times as a parameter to calculate fuel consumption. Therefore, the improvements in number of holds would not have been considered at all. The same goes for the ICAO approach of using the fuel consumption at idle thrust settings, which also only considers taxi time. However, the lower number of average holds with only a minimal increase in taxi time is a clear indication that the green profile is more efficient and can contribute to more environmental-friendly taxi operations. Specifically, the technology used to achieve this requires no additional investment in airport infrastructure or a change in traffic structure, but is achievable by implementing software systems only. Therefore, this kind of optimization can be a fast way to improve the environmental impact of taxi operations.

The assessment of the fuel efficiency and environmental sustainability of traffic in the TMA proved more challenging. It was only possible to assess main operational efficiency for the comparison of the HITL trials with the reference scenarios, since the reference scenario data lacks data points to conduct a detailed fuel estimation, that takes the different descent profiles into account. However, the analysis of flown track miles during the approach phase showed, that the different scenarios with 4D-FMS equipped aircraft overall had a lower average number of track miles flown by the aircraft compared to the reference scenario based on real traffic data.

A detailed assessment of the fuel efficiency results and environmental sustainability for the traffic in the TMA for the HITL simulation runs has been conducted by UPM, indicating clear benefits for the 4D-FMS approaches [Alonso 2023].

EXE-002 – HC & PILDO LABS

Minor improvements have been observed in terms of mean fuel consumption and CO_2 emissions.

Table 34. Measured benefits of MergeStrip (per flight).

Indicator	MergeStrip	No MergeStrip	Benefit
Fuel consumption	361.26 kg	364.24 kg	-2.98 kg
CO_2 emissions	1137.93 kg	1147.36 kg	-9.43 kg

These results obtained during the shadow-mode validation exercise have to be put into context, as there were several factors influencing the exercise which were beyond the control of project partners.

It has to be mentioned that with the redesign of Budapest TMA entering into effect in January 2020, the local maximum level of efficiency has been achieved. In this respect, the tested tool was able to make only minor improvements. Furthermore, Budapest TMA traffic is still below the pre-COVID level, and also below what was expected for 2023. Considering that MergeStrip is expected to bring most benefits under high traffic scenarios, this fact had a negative impact on the measurements (its potential in high traffic scenarios could not be assessed). Finally, as a consequence of the war in Ukraine, the number and occurrence of TRAs have increased significantly, and these TRAs hinder aircrafts to fly the optimal vertical profile. In this context, even this minor improvement development can be considered a very important one.

4.3.4. UNEXPECTED BEHAVIOURS/RESULTS

No unexpected behaviours were reported or noticed during the trial EXE-001.

During EXE-002, The instability of the system in the early phases of the validation exercise was unexpected and had an effect on the general impression of the ATCOs. This instability issue was resolved quickly however it left its mark on the later phases as well, i.e. in the forms of answers.

4.4. CONFIDENCE IN RESULTS OF THE VALIDATION EXERCISES

4.4.1. QUALITY OF VALIDATION EXERCISES RESULTS

- **EXE-001 – DLR**

Data collection and data analysis were appropriately monitored and are assumed to be of good quality. The system ensured that the answers were complete and one member of the validation team was present at all times to answer questions. Participants completed the questionnaires post-run and post-exercise, see Section 3.1.5. Questionnaires and debriefing sessions alike were scheduled and carried out adequately in order to capture the ATCOs' experiences.

Results regarding training effects and simulation quality are reported in ANNEX 7.2 The simulations were experienced as sufficiently realistic. Training seemed to be sufficient, but training effects over the course of the day cannot be ruled out.

- **EXE-002 – HC & PILDO LABS**

Real-time Simulation

The quality of the results is in line with what is reported in the above paragraph for EXE-001. The participant team consisted of six licensed air traffic controllers and 2 validation leaders who are also licensed APP ATCOs. The participating ATCOs have filled out the questionnaires and added their thoughts as text as well, and participated on the debriefing sessions proactively. All their needs and ideas are outlined in Section 4.3 in great detail and the system has been further developed by considering these points.

Shadow Mode

The results have been obtained by four major sources: two types of questionnaires, observation and a final debriefing discussion, analysing the outcomes of the last questionnaire. Therefore, the results represent the true thoughts and feelings of the participating ATCOs. The number of participants speaks also highly of the volume of this validation. Nevertheless, the quality and depth of the human performance results match the quality of the system state that was tested.

4.4.2. SIGNIFICANCE OF VALIDATION EXERCISES RESULTS

- **EXE-001 – DLR**
  - There was no baseline scenario to compare the human performance results against.
  - Explanatory power is limited due to the small sample size of N = 5 per iteration.
  - The human performance data were analysed on a non-parametric, descriptive level only, i.e. no statements can be made regarding statistically significant differences.
  - Because the order of simulation runs was held constant for all participants, training effects or effects of exhaustion cannot be ruled out. Thus, potential differences in human performance between simulation runs cannot be attributed to the traffic distribution.
- **EXE-002 – HC & PILDO LABS**

Real-time Simulation

- In the first iteration, there was baseline scenario to compare the human performance results against, but there were no significant differences between the median workload and situational awareness values.
- Explanatory power is limited due to the small sample size of N=8.
- The human performance data in the second iteration were analysed on a non-parametric, descriptive level only, i.e. no statements can be made regarding statistically significant differences.
- The scenario order has been changed for the two groups of participants sitting in different circuits in order to minimise the scenario order effect. For instance, one group started with scenario 105 and ended with 204, and the other with 106 and ended with 203.

Shadow mode

- In light of the results the validation team gained an in-depth knowledge about what an advanced APP decision support tool can be capable of.
- More robust system is needed for validation of a web-based tool having ML based functionalities. Once the stability issues were solved, the system functioned properly. However, this degraded the users’ confidence in the new version, and this first impression could not have been overwritten even by swift and effective corrective measures.
- Explanatory power is limited due to the small sample size of N=11 (approximately 25% of APP staff).

Positive externality: BUDAPEST TMA was well redesigned, and there is small room for improvement in terms of fuel efficiency under the current traffic size and pattern.

5. CONCLUSIONS

This document provides the Validation Report for GreAT project. It summarizes the short haul flight validation exercises for the new GreAT airspace, arrival manager and display support functions and Merge Strip, all previously defined in the validation plan [Kling 2021]. It describes how they have been conducted and provides analyses, conclusions, recommendation as well as potential next steps. The Validation Report conveys the overall series of validation activities with the aim of delivering results that may contribute to the successful implementation of GreAT concept elements for short haul flights to reduce the emissions in the future.

In GreAT project, the goal was to develop techniques and procedures to reduce the environmental impact of aviation. In the project, a distinction was made between long haul and short haul flights. This document now described the results from the validations obtained for short haul operations.

The validation phase included two main focus areas. The first was to show that a technical implementation of the support functions for different pilot workstations is possible. Secondly, it was to be shown that controllers can use it to work at their workstations, thus making it possible to reduce pollution for the environment without affecting safety and without causing major capacity losses at an airport or in an airspace. Both automatic and human-in-the-loop simulations were used as validation techniques. Data collection was both automatic for extensive statistical analysis and descriptive through interviews with the test participants, all of whom were from professional air traffic control environments.

By redesigning the airspace in the area of today's TMAs and the use of a route and target time negotiation between board and ground, it became possible to separate aircraft with different levels of technical equipment along different routes, so that a large proportion can perform their own optimized approaches. In addition, a free space was created for the approaches along an optimized approach profile for the most direct approach routes possible to the final. In this way, approach distances could be reduced by an average of six Nautical miles. The departure routes, on the other hand, were considered, but it became apparent that these would have to be further optimized in order not to lose some of the benefits of the new approach routing. The new airspace has some special features. For example, there are more crossing points between the approaches, since each approach must be performed from each cardinal direction. The intercept to the final is therefore at 8000 ft. as opposed to the 4000 ft. or 5000 ft. usual today. In order to still have enough distance to reduce altitude and speed, the final was extended to 25 nautical miles. However, validations subsequently showed that this would not have been necessary with scheduling AMAN support.

Controllers were provided with different visual and planning support systems for the new airspace and the challenge of merging two approach flows on final. It turned out that it was no problem for the controllers to use these and thus ensure safe and efficient approach guidance. The trajectory-based optical support methods Ghosting and TargetWindows performed best, the additional window with the Final Distance Indicator for precise distance detection on the final proved to be superfluous, since a visually appealing scale on both the final and the downwind performs this task just as well.

In the validations, a maximum workload was simulated by requiring controllers to operate up to three workstations simultaneously under a normal traffic mix. The results showed very clearly that the AMAN support in combination with the negotiated approaches enabled the controllers to handle the traffic in a safe and focused manner.

The generation of more environmentally friendly taxi trajectories was the scope of the improvement of SMAN software. By the eliminating of unnecessary holding times with the help of conflict resolution algorithms in combination with a parameterized optimization function with higher penalties for holds during normal taxi, taxiing profile calculation were configured to prefer holding aircraft at their stand before engine start-up. These measurements resulted in a reduction of the number of holds by 80% in high-density traffic scenarios in comparison to a conventional trajectory profile. Furthermore, the green profile consistently produced trajectories with fewer holds than usually scheduled today, regardless of traffic density and palnning times.

The most important result of the MergeStrip developments for Budapest airport and their validation is that it did help reducing the carbon footprint of TMA operation. At the same time, it worth putting it into context as there were several factors influencing this shadow-mode validation, and which were beyond the control of project partners.

- It has to be mentioned that with the redesign of Budapest TMA entering into effect in January 2020, the local maximum level of efficiency has been achieved. In this respect, the tested tool was able to make only minor improvements.
- Furthermore, given the fact, that Budapest TMA traffic is still below the pre-COVID level, and also below what was expected for 2023. Finally, as a consequence of the war in Ukraine, the number and occurrence of TRAs have increased significantly, and these TRAs hinder aircrafts to fly the optimal vertical profile, especially from runway direction 31 (south east arrival flow).
- As environment was the key focus, it has to be mentioned that even against these very unfavourable outside circumstances, even this minor improvement in terms of fuel savings can be considered a very important one.

Specific features, improvement over the previous version that shall be kept for further development:

- Test New Speed worked well, as position change can be seen at once
- VPV shows the one-minute vectors as well → rate of descent appears at once
- Automatic detection of Arrivals is a huge help in high traffic
- „Distance to previous” seems more accurate than in the current version
- Does not only calculate with the current speed: good point
- Bigger radar coverage than the current one
- Calculation with all waypoints of the STARs
- Enabled handling 7-8 aircrafts (vs 5-7 currently)

As a conclusion, it might be stated that the developments on TRL-4 level (as prescribed by the Call for Proposals) overall proved well. Regarding future prospects, the main directions for improvement could be i) the inclusion of wind and ii) the application of another data source besides ADS-B. Crucial factor for stepping forward is the guidance from EASA and more importantly regulation on AI/ML in ATM. Without these pieces of legislation, any further development might be in vain, as no permit can be obtained from any National Supervisory Authority for putting the into operation.

The results of the short haul validations show on the one hand that the existing systems and processes for approach guidance are already working in many areas close to the current technical optimum. Improvements per approach are only possible gradually and amount to around 40-80 litre per approach. However, extrapolated to more than 11 million approaches in 2019 in Europe alone, this results in at least 440,000 tons of kerosene or correspondingly almost 1.4 million tons of CO_2. All in all, however, it is clear that the potential for reducing climate-damaging emissions exists and that this potential could be exploited by separating traffic flows and additional air traffic controller support. However, the exact magnitude of this potential depends on so many local factors that general statements on the climate effectiveness of individual measures can only be made to a very limited extent.

6. REFERENCES

[Åkerstedt 1990] Åkerstedt, T., & M. Gillberg. (1990). Subjective and objective sleepiness in the active individual," International Journal of Neuroscience, vol. 52, no. 1-2: pp. 29-37.

[Alonso 2023] Alonso, G., A. Benito and À. Ramonjoan (2023). Environmental impact assessment and green trajectory selection. GreAT D7.3. Universidad Politécnica de Madrid. Madrid, Spain.

[Athènes 2002] Athènes, S., P. Averty, S. Puechmorel, D. Delahaye and C. Collet (2002). ATC Complexity and Controller Workload: Trying to Bridge the Gap. HCI-Aero, American Association for Artificial Intelligence (AAAI). Cambridge, Massachusetts, USA.

[Bangor 2009] Bangor, A., P. Kortum and J. Miller (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies, vol. 4, no. 3: pp. 114-123.

[Brochard 2019] Brochard, M. and J. Engel (2019). PJ19 CI: Final Project Report. D1.2. EUROCONTROL Exp. Centre. Bretigny sur Orge, France.

[Brooke 1996] Brooke, J. (1996). SUS - A Quick and Dirty Usability Scale. Usability Evaluation In Industry, P. W. Jordan, B. Thomas, M. I. L. und B. Weerdmeester, Ed., London, Taylor & Francis Ltd.: pp. 189-194.

[Dehn 2008a] Dehn, D. M. (2008). Assessing the impact of automation on the air traffic controller: the SHAPE questionnaires. Air Traffic Control Quarterly, vol. 16, no. 2, pp. 127-146.

[Dehn 2008b] Dehn, D. M. (2008). Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires. Air Traffic Control Quarterly 16(2): pp. 127-146.

[DIN ISO 9241 2020] DIN EN ISO 9241-210:2020-03. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems (ISO 9241-210:2019); German version EN ISO 9241-210:2019.

[Endsley 1995] Endsley, M. R. (1995). "Toward a Theory of Situation Awareness in Dynamic Systems." Human Factors: The Journal of the Human Factors and Ergonomics Society 37(1): pp. 32–64.

[EUROCONTROL 2010] EUROCONTROL (2010). European Operational Concept Validation Methodology - E-OCVM. Version 3.0, Volume 1. EUROCONTROL and European Commission, Brussels, Belgium: EUROCONTROL.

[EUROCONTROL 2012] EUROCONTROL (2012). SATI - SHAPE Automation Trust Index. [Online]. Available: https://ext.eurocontrol.int/ehp/?q=node/1594. [Accessed 07 06 2023].

[Finke 2021] Finke, M., M.-M. Temme, R. Abdellaoui, E. Zoltán, G. Csaba, M. Gábor, P. A. Barna, T. Péter, S. Péter, H. Haoliang, L. Guang, Y. Peng and Y. Lei (2021). GreAT D2.1: Current TBO Concepts and Derivation of the Green Air Traffic Management Concepts. German Aerospace Center (DLR). Braunschweig, Germany.

[Grier 2015] Grier, R. A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting. Los Angeles, California, USA.

[Hamann 2020] Hamann, A. and N. Carstengerdes (2020). Fatigue Instantaneous Self-Assessment (F-ISA) - Development of a Short Mental Fatigue Rating. DLR-IB-FL-BS-2020-64. German Aerospace Center (DLR), Braunschweig, Germany.

[Hamann 2022] Hamann, A. and N. Carstengerdes (2022). Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights. Scientific Reports, vol. 12, no. 1, p. 6449.

[Hart 1988] Hart, S. G. and L. E. Staveland (1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." Advances in Psychology 52: pp. 139-183.

[Hart 2006] Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting.

[IBM 2019] IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0, Armonk, New York: IBM Corp.

[ICAO 2005] ICAO (2005). Global Air Traffic Management Operational Concept. Doc 9854 AN/458. International Civil Aviation Organization (ICAO).

[Kling 2021] Kling, F., Z. Molnár, B. Nagy, A. Pásztor, T. Mühlhausen, M.-M. Temme and Á. Ramonjoan (2021). GreAT D6.1 Validation Plan. HungaroControl, Budapest, Hungary.

[Nuic 2010] Nuic, A., D. Poles and V. Mouillet (2010). "BADA: An advanced aircraft performance model for present and future ATM systems. International Journal of Adaptive Control and Signal Processing 24: 850–866.

[OpenSky 2023] OpenSky (2023). A Quick Guide To OpenSky's Impala Shell. OpenSky Network. Burgdorf, Switzerland.

[Tattersall 1996] Tattersall, A. J. and P. S. Foord (1996). An Experimental Evaluation of Instantaneous Self-Assessment as a Measure of Workload. Ergonomics, vol. 39, no. 5, pp. 740-748.

[Temme 2021] Temme, M.-M., M. Kleinert, R. Abdellaoui, M. Finke, I. Gerdes, M. Schaper, F. Kling, B. Nagy, T. Boldogh, Z. Eszes, Z. Molnár, A. B. Pásztor, H. Haoliang, O. Ohneiser and A. Ramonjoan (2021). Environmental-friendly airspace structuring and traffic sequencing. D4.1. German Aerospace Center (DLR). Braunschweig, Germany.

7. ANNEX

7.1. PROPOSED IMPROVEMENTS (HUM–GREAT–07)

The following section lists improvements regarding the usability of the system that were proposed by the ATCOs during EXE-001. The comments are arranged by criterion ID. It is recognized that, in the case of usability, the mapping of comments to one of the three criteria is not always unambiguous and thus some comments could be mapped to several attributes of the usability.

7.1.1. HUM-GREAT-CRT-07-10

HUM-GREAT-CRT-07-10 - There is no discrepancy between system-provided information and user-required information.

Table 35. Proposed improvements regarding HUM-GREAT-CRT-07-10.

Feature	Proposed improvement
First Iteration	
General	Visual load should be reduced
AMAN	The AMAN should provide several options (straight calculation to LMP and also different calculations to other LMP)
AMAN	The AMAN should consider departures
Ghost	Reduce the amount of information displayed in the ghost' label
Ghost	There should be more options for the ATCO to customize the ghosts
Target Window	Reduce amount of lines in target window (used more compact way to display the window)
Target Window	Reduce the amount of information displayed in target window label
Target Window	Display the target window on demand: mouseover function
Target Window	Display the target window on demand: button to switch on/off
Target Window	Additional feature: A line that indicates when to turn
Target Window	Additional feature: A Countdown when to turn to hit the target window
AMAN	When a target window is missed, there should be a recalculation of the aircraft sequence and the window should be updated accordingly
Second Iteration	
Radar screen	Display the Flight leg for a selected aircraft
Radar screen	Feature to compute the "track to go distance"
AMAN	Should be able to recognize shortcuts and adapt the sequence accordingly
AMAN	To be able to overwrite the AMAN sequence
Radar screen	provide a feature to enable measuring the distance between ghosts and aircraft
Ghost	Display the ground speed on final

Feature	Proposed improvement
Radar screen	Additional feature to check the correctness of the ghost position
Ghost	Ghosts should disappear earlier
Ghost	Ghosts should appear later
Ghost	It should be possible to turn ghost labels on and off
Ghost	Reduce information in the ghost labels (avoid overlapping)
Ghost	More information should be displayed in the ghost labels, e.g. vertical speed, ground speed
Target Window	The target window should be smaller/ more precise: Only show circle and a buffer (1 mile/side) or a circle and two lines before and after
Target Window	Display the target window on demand
AMAN	Provide an advisory list of clearances/ steps to be implemented in order to precisely guide the aircraft to the optimum position on the target window.
Target Window	Reduce the amount of information displayed in the target windows' label
Target Window	The target window should appear later
Final Distance Indicator	Display future distances with the current speeds
Final Distance Indicator	Display projected distances to the threshold (extrapolated distance between aircraft by the time the first aircraft gets to the threshold; shows separation loss)

7.1.2. HUM-GREAT-CRT-07-20

HUM-GREAT-CRT-07-20 - The ATCO can perform interaction without noticeable problems

Table 36. Proposed improvements and challenges regarding HUM-GREAT-CRT-07-20.

Feature	Proposed improvement/Challenge
First iteration	
Radar screen	Challenge: It was troublesome to make input in the interface
Radar screen	Challenge: Handling labels and distances in the new airspace
Concept	Challenge: The TMA was too big (need to zoom in/ out)
Radar screen	Challenge: Assuming and transferring aircraft
Radar Screen	The menu of the speed vectors should be a scroll down menu
Second Iteration	
Ghost	Ghost labels should be moveable (not cover the final scale)
Target Window	Should be more stabilized (avoid jumping, disappearing, etc.)

7.1.3. HUM–GREAT–CRT-05-40

HUM – GREAT – CRT-05-40 - The look-and-feel of the HMI is acceptable.

Table 37. Proposed improvements regarding HUM-GREAT-CRT-05-40.

Feature	Proposed improvement
First Iteration	
Radar screen	Checked-in aircraft should have a different color
Ghost	The discriminability between ghosts and aircraft should be improved
Ghost	Ghost labels should be smaller
Ghost	Ghost labels should be moveable (to avoid overlapping and covering the final)
Ghost	The shape of the ghosts should be changed (such type of symbol is normally used to indicate the type of surveillance data used to compute the current position of aircraft)
Ghost	Ghosts should not be filled out
Ghost	Change shape/color/position of labels in the ghosts
Target Window	The display of the target windows should be changed (only showing boarders)
Target Window	The label should be better positioned (not too far outside the window)
Target Window	Reduce amount of lines in target window
Target Window	Add color to target window to make it stand out
Second Iteration	
Radar screen	Use color coding to mark aircraft that are the ATCO’s responsibility (e. g. mark assumed aircraft brighter)
Ghost	The display of the ghosts should be brighter to improve detectability
Ghost	The discriminability between ghosts and aircraft should be improved
Ghost	Visual load should be reduced (not too dense)
Target Window	Add colour to target window
Target Window	The target window should be narrower
Final distance indicator	The provided information should be displayed in better location in the screen (not on the bottom, rather more directly where the traffic converges/ is located)

7.2. RESULTS CONCERNING TRAINING EFFECTS AND QUALITY OF SIMULATION

In order to check for training effects and to evaluate the quality of the simulation, participants were asked to rate the following two statements post-training and post-run:

1. I feel well acquainted with the simulation.
2. I felt immersed in the simulation.

The statements were rated from 1 (strongly disagree) to 5 (strongly agree) and mean ratings were calculated. Mean ratings of 3 or above for (1) were interpreted as a sufficient level of training. Mean ratings of 3 or above for (2) were interpreted as the simulation being of sufficient quality.

FIRST ITERATION

Figure 82 shows the mean agreement to the statements regarding the simulation depending on the run number. On average, participants gave a rating between 3 (neither agree nor disagree) to 4 (agree) to the statement "I feel well acquainted with the simulation" for each run, pointing towards a successful training. The average agreement to the statement "I felt immersed in the simulation" was between 3 (neither agree nor disagree) and 4 (agree) for each run, indicating a sufficient simulation quality.

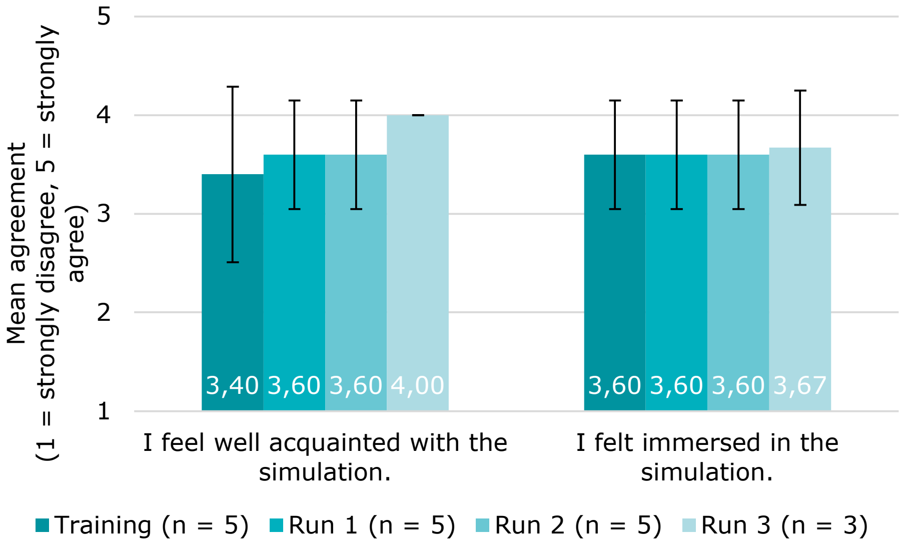


Figure 82. Mean agreement to questions regarding the simulation by number of run (training vs. run 1 vs. run 2 vs. run 3). Error bars represent standard deviations.

The following question was asked during the debriefing:

How realistic was the simulation environment?

- Three ATCOs reported that the realism of the simulation was sufficient
- One ATCO stated that the realism of the simulation was good besides the position of the LMP
- One ATCO had no answer to this question

SECOND ITERATION

Figure 83 shows the mean agreement to the statements about the simulation depending on the run number. It can be seen that the agreement to "I feel well acquainted with the simulation" increased from the training run to the first simulation run to the second simulation run. This could point to a training effect. Mean ratings for this statement ranged between 3 (neither agree nor disagree) and 5 (strongly agree), pointing towards a sufficient level of training. Regarding the immersion in the simulation, participants gave mean ratings around 4 (agree) for all runs. This was interpreted as the simulation being of sufficient quality.

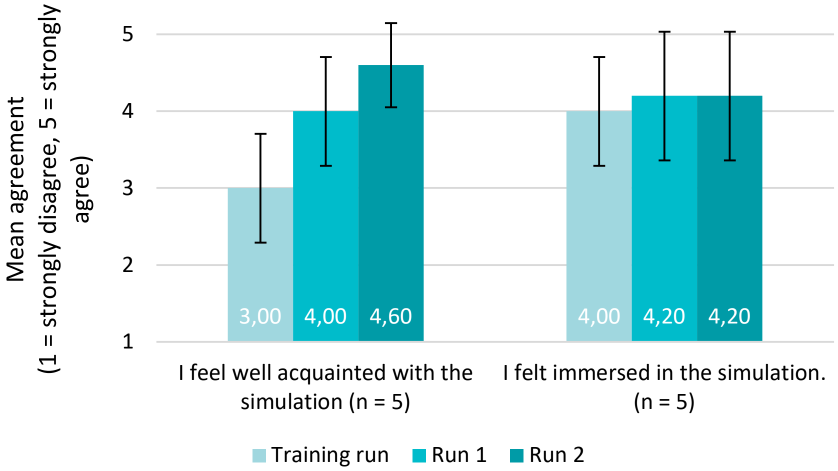


Figure 83. Mean agreement to questions regarding the simulation in dependence of number of run (training vs. run 1 vs. run 2). Error bars represent standard deviations.

The following question was asked in the final tailored questionnaire:

Were there elements in the simulation that seemed unrealistic to you?	Number of answers
Yes	1
No	4

The ATCO affirming this question stated that the speed of aircraft close to the threshold seemed higher to him compared to real life. Another ATCO expressed during the debriefing that he perceived the departures as unrealistic.