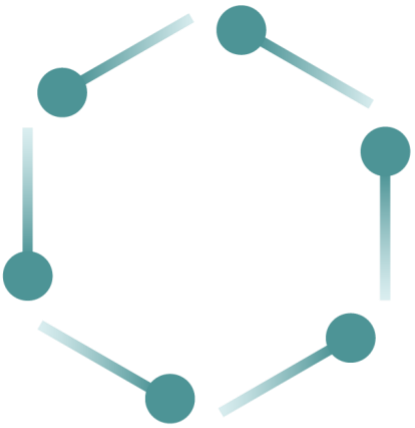
Ref. Ares(2025)5215836 - 30/06/2025



SafeTeam

D4.2 - Human-machine collaboration in destabilized approaches

Document number	D4.2
Document title	Human-machine collaboration in destabilised approaches
Version	1.0
Work package	WP4
Edition date	30.06.2025
Responsible unit	Technical University of Munich
Dissemination level	PU
Project acronym	SafeTeam
Grant	101069877
Call	Safe, Resilient Transport and Smart Mobility services for passengers and goods (HORIZON-CL5-2021-D6-01)
Topic	HORIZON-CL5-2021-D6-01-13: Safe automation and human factors in aviation – intelligent integration and assistance

This project has been funded by the European Union under Grant Agreement 101069877

Funded by the European Union

© SafeTeam Consortium.

SafeTeam Consortium

	Innaxis (INX)
	Agencia Estatal de Seguridad Aérea (AESA)
	Technische Universität München (TUM)
	DataBeacon
	ONERA
	Rise Research Institutes of Sweden AB (RISE)
	PEGASUS HAVA TASIMACILIGI ANONIM SIRKETI (PEGASUS)
	CAA INTERNATIONAL LIMITED (CAAi)

Document change record

Version	Date	Status
0.1	15.06.2025	Draft
1.0	30.06.2025	Initial Submission
1.1	30.06.2025	Final version for submission

Abstract

In work package 4.2 of the SafeTEAM project, a Stabilized Approach Digital Assistant, which was conceptualized in Deliverable 3.2 of SafeTEAM, is integrated into the research simulator at TUM. In simulation exercises with several airline pilots, the Human-Machine Collaboration of the Digital Assistant is tested. To design the simulator exercise, this deliverable applies the performance monitoring framework of WP2 with predefined performance metrics to assess the impact the digital assistant has on its end users. For this, scenarios will be defined that measure human performance and the digital assistant's influence on safety and resilience.
The results substantiate the digital assistant's benefit on safety, especially regarding go-around compliance, while also providing evidence-based hits on further improvement regarding the computation and presentation of the assistant's advisories.

Disclaimer: To improve the spelling and grammar of this document, we use Grammarly, an AI-based writing assistant. The AI is not used to generate any content but solely to improve the reading experience, based on content produced by the authors.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ATC	Air Traffic Control
FDM	Flight Data Monitoring
ft	feet (Unit)
HMI	Human Machine Interface
ILS	Instrument Landing System
LLM	Large Language Model
NASA TLX	NASA Task Load Index
PFD	Primary Flight Display
SADA	Stabilized Approach Digital Assistant
SA	Situation Awareness
TA	Thematic Analysis

1 Executive Summary

The SafeTeam project elaborates human factors related to AI integration through several practical use cases, in which we aim to progress the safe introduction of automation in the form of intelligent assistance to humans.

This deliverable describes the planning, conduction and evaluation of a simulator case study for one of these use cases of SafeTEAM, the Stabilized Approach Digital Assistant (SADA). It continues the work from Deliverable 3.2 (SafeTEAM) of SafeTEAM, which describes the design and concept of the Stabilized Approach Digital Assistant, based on an Unstable Approach Prediction Model (Martinez, et al., 2019), which in turn is an artifact of the SafeClouds.eu (SafeClouds Consortium, kein Datum) project.

The first part of this deliverable, the Experimental Plan, describes the implementation of the SADA concept into a research simulator at the Technical University of Munich's Institute of Flight System Dynamics. Thereby, SafeTEAM raised the TRL of the SADA from 4 to 6, by integrating the Unstable Approach Prediction Model as a constituent in the SADA and demonstrating the SADA in a relevant environment. The description of the implementation comprises a summary of Unstable Approach Prediction Model, the high-level layout of the simulator, including a description of the aircraft model, the relevant cockpit human machine interfaces as well as the interfaces between these components.
Based on SADA's implementation, the experimental plan contains the posed research questions. These questions aim at the potential safety benefits of the SADA, specifically the go-around compliance of pilots during unstable approaches, as well as the potential to prevent avoidable unstable approaches, which in turn could increase the resilience of the aviation system, especially the landing phase.

To collect the relevant information with respect to the research questions posed, the Experimental Plan defines a set of test scenarios as well as the metrics to evaluate the simulator exercises. The experimental plan is a combined effort of the SafeTEAM work packages 2 and 4. On the one hand, the experimental plan applies SafeTEAM's human performance monitoring framework, developed in work package 2, to the use case of the SADA. On the other hand, the gathered experiences were fed back to the framework development on a continuous basis.

The second part of this deliverable contains the Experiment Evaluation. Work on the data collection and evaluation started only after the experimental plan was compiled.
Based on the collected qualitative and quantitative data, the evaluation substantiates the potential benefits of the SADA concept for aviation safety, especially with respect to go-around compliance of pilots in unstable approach situations. Furthermore, it shows that the chosen Human Machine Interfacing (HMI) concept is overall well-received by pilots, with minor modifications requested by a few participants.
The simulation exercise, however, also showed that the SADA, in its currently implemented form, does not fully meet the objective of preventing avoidable go-arounds, in the case of high-speed approaches. Nevertheless, the exercise provides an evidence-based path on necessary modifications and future work, especially regarding the Unstable Approach Prediction Model, to achieve the not yet completely fulfilled objective.

2 Introduction

Unstable approaches are one important precursor for accidents related to the approach and landing phase. According to the Flight Safety Foundation, seven fatal accidents of commercial aircraft between 2009 and 2013 claimed 191 lives as the aircraft ran off the runway following an unstabilized approach (International Air Transport Association, International Air Transport Association 2014 – Runway Excursion Statistics, 2014).
The Flight Safety Foundation Approach-and-Landing Accident Reduction task force found that unstabilized approaches were a causal factor in 66% of 76 approach-and-landing accidents and serious incidents worldwide in 1984 through 1997. The task force found that some low-energy approaches caused loss of aircraft control and involvement in controlled flight into terrain. High-energy approaches caused aircraft loss of control, too, resulting in runway overruns and runway excursions, and contributed to inadequate situational awareness. The causal factor in 45% of the 76 approach-and-landing accidents and serious incidents was the crew's inability to control flight parameters such as the airspeed, altitude, and rate of descent. These flight-handling difficulties originated from rushing approaches, attempts to adhere to demanding ATC clearances, adverse wind conditions, and improper use of automation (Flight Safety Foundation, 2000).

This deliverable describes the simulator testing of a Stabilized Approach Digital Assistant (SADA). The concept for this digital assistant was developed in Deliverable 3.2 of SafeTEAM (SafeTEAM), based on a user-driven approach. The core idea of the concept is to provide pilots with an indication before reaching the stabilization gate, if their approach is prone to becoming unstable, and by this increase aviation safety with respect to the stated approach and landing accidents. This deliverable is a continuation of the work started in Deliverable 3.2, and while it provides a summary of the SADA concept necessary to follow this document, it does not describe the concept at the level of detail but focuses on the testing of the concept in a simulator environment. The remainder of this document is structured as follows.

This rest of this section provides more information on unstable approaches, mainly how they are defined in the industry and for this project.
The first part of this deliverable, following this introduction, is an Experimental Plan, which describes:

- the Stabilized Approach Digital Assistant (SADA) as currently implemented in the research simulator, in section 3,
- the research questions this deliverable poses, in section 4,
- as well as the methodology to gather the information needed to evaluate the posed research questions, section 5.

The Experimental Plan was finalized before conducting the experiments. The second part is the Experiment Evaluation, which covers:

- The results of the experiments, in section 6,
- The discussion of the results, necessary modifications to the SADA and potential next steps in section 7.

2.1 Unstable Approach Definitions

The International Civil Aviation Organization (ICAO) states a recommendation for operators to define stabilized approach procedures in Annex 6 Part 1 (2.1.25) (International Civil Aviation Organization 2022 – Annex 6: Operation of Aircraft, 2022). ICAO does not specify stabilized approach procedures itself, but requests operators to define them in their operations manual. Therefore, stabilized approach procedures between operators might differ slightly. In the following, we provide an overview of different stabilized approach definitions by various stakeholders and finally define the definition of an unstable approach for this exercise. This way, we can ensure that all pilots taking part in this exercise, even though they work for different operators, have a common definition for the simulation exercises.

2.1.1 Flight Safety Foundation

In its Approach and Landing Accident Reduction briefing note 7.1 (Flight Safety Foundation, 2000), the Flight Safety Foundation recommended minimum stabilization heights at 1000 ft above aerodrome elevation in instrument meteorological conditions and 500 ft above aerodrome elevation in visual meteorological conditions. An approach is considered stabilized when these parameters are met:

1) The aircraft is on the correct flight path according to navigation aids or visually.
2) Only minor changes in heading/pitch are required to maintain the correct flight path.
3) The aircraft speed is not more than the reference speed (VREF) + 20 knots indicated airspeed (IAS) and not less than VREF.
4) The aircraft is in the correct landing configuration.
5) The rate of descent is no greater than 1000 feet per minute; if an approach requires a sink rate greater than 1000 feet per minute, a special briefing should be conducted before.
6) The power setting is appropriate for the aircraft configuration and is not below the minimum power for approach as defined by the aircraft operating manual.
7) All briefings and checklists are conducted.
8) Specific types of approaches are stabilized if they also fulfill the following:
    a) Instrument landing system (ILS) approaches must be flown within one dot of the glide slope and localizer.
    b) A category II or category III approach must be flown within the expanded localizer band.
    c) During a circling approach, wings should be level on final when the aircraft reaches 300 feet above airport elevation
9) Unique approach procedures or abnormal conditions requiring a deviation from the above elements of a stabilized approach require a special briefing.

An approach that doesn't meet one or more elements of the above-mentioned criteria is considered unstable - below 1000 feet above airport elevation in IMC or below 500 feet above airport elevation in VMC - and requires an immediate go-around [1].

2.1.2 International Air Transport Association

The International Air Transport Association endorses the criteria for a stable approach established by the Flight Safety Foundation Approach and Landing Accident Reduction briefing note 7.1. Therefore, an approach is assumed to be unstable if not all these parameters are met at the stabilization gate, which is at 1000 ft above airfield elevation in instrument

meteorological conditions or at 500 ft above airfield elevation in visual meteorological conditions.
In the Threat & Error Management terminology, an unstable approach is an undesired aircraft state, which the flight crew can recover to prevent an unrecoverable outcome (accident). A perceived unstable approach can be managed using established recovery techniques to prevent accidents. Eventually, the pilot must execute a go-around if the aircraft is unstable with respect to the stabilized approach criteria mentioned above. If carried out properly, the go-around maneuver is considered the safest course of action. Not going around, however, has been identified as a contributing factor in approach and landing accidents (International Air Transport Association, International Air Transport Association 2017 – Unstable Approaches, 2017).

2.1.3 European Aviation Safety Agency and Data4Safety

For operators, the European Union Commission Regulation (EU) No. 965/2012 (European Comission, 2012-10-05) provides guidelines for flight data monitoring (FDM) pilot training and recurrent assessment. However, no thresholds, gates, or parameters are specified. As the EASA-Data4Safety Programme evolves, more specific parameters are defined to guide flight data analysts to identify unstable approaches. The initiative aims for a complete characterization of the logic for the detection of unstable approach events, present the different criteria and thresholds that the identification of instabilities encompasses to guide industry practitioners on its implementation, and convey a set of assumptions, considerations, and lessons learned, arising out of the work performed during the definition of unstable approach detection algorithm, aiming to assist industry practitioners when conducting safety analysis in this area (Data4Safety, 2022), which is summarized in the remainder of this subsection.
Data4Safety defines an unstable approach from the perspective of the flight data analysis as "any approach with the minimum required instability conditions triggered within the analysis window (1000 ft– oft) as per the instability criteria and height band".
Commonly, instability conditions to be analyzed include:

- Approach speed above/below the desired reference speed
- Vertical speed too high
- Aircraft misconfiguration (landing gear or flaps)
- Engine thrust level
- Approach path deviations

This leads to instability criteria within the approach window (<1000 ft):

- Fast descent
- Low thrust
- High/Low airspeed
- TAWS alert
- Late flap or gear extension
- Unstable attitude (pitch/roll)
- High/low glide slope and localizer deviation

Each criterion is evaluated in different height ranges and thresholds to assign a severity level. A distinction is made between the two ranges above/below 500 feet. An approach is considered unstable if a minimum number of criteria are met. Above 500 feet, three criteria

must be met, while below this threshold, only one instability condition triggers the classification as an unstable approach.

2.1.4 SafeTEAM Simulation Exercise Definition

To take into account the variations in the definitions for unstable approaches, the basic parameters of the Flight Safety Foundation for airline operations are used in the simulator exercises within SafeTeam.
These parameters include:

- minimum stabilization height 1000 ft above airfield elevation.
- The aircraft is on the correct flight path according to navigation aids or visual.
- Only minor changes in heading/pitch are required to maintain the correct flight path.
- The aircraft speed is not more than VREF + 10 knots indicated airspeed (IAS) and not less than VREF -5 kts.
- The aircraft is in the correct landing configuration.
- The rate of descent is no greater than 1000 feet per minute.
- The power setting is appropriate for the aircraft configuration and is not below the minimum power for approach as defined by the aircraft operating manual.
- All briefings and checklists are conducted.

An approach that doesn't meet one or more elements of the above-mentioned criteria is considered unstable and requires an immediate go-around.

Experimental Plan

3 Solution to be Evaluated

This section of the Experimental Plan summarizes the Digital assistant under investigation. It first describes the underlying machine learning model and the Primary Flight Display (PFD) modifications to visualize the prediction results to pilots.

3.1 Simulator

The research simulator used for this study is a self-built and flexible simulator that builds around a generic cockpit design, which is oriented on an A320 cockpit. The aircraft model which is simulated is a high-fidelity model of a Dornier 728 jet, a two engine turbo jet aircraft that was developed by Dornier Fairchild. Since the simulator is entirely custom built, it allows for easy adaptation of the PFDs and also easy implementation of a machine learning model, since the software for the displays and the flight dynamics model is developed in-house. Figure 1 illustrates the aircraft, rendered by the visualization software used in the simulator.

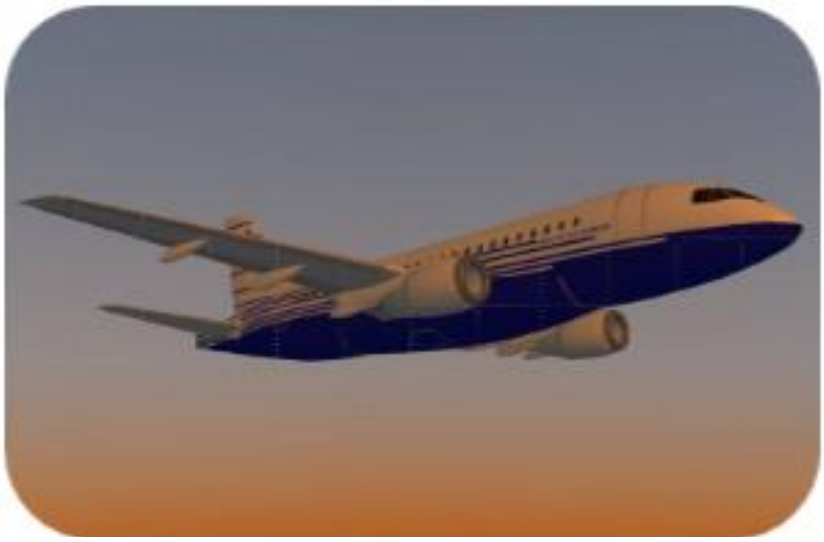


Figure 1: Rendering of the Do-728 Jet in the simulator visualization

The following subsections describe the implementation of the SADA, as outlined in Deliverable 3.2 of SafeTEAM, in more detail.

3.2 Implementation of the Stabilized Approach Digital Assistant

Figure 2 illustrates the implementation of the SADA in the Research Simulator, with all relevant simulator modules. Each grey box indicates a computer, whereas the blue boxes illustrate the simulator modules (software). The Flight Dynamics Model, Flight Controls, PFDs, and Instructor Station are existing parts of the research simulator. The Unstable Approach Prediction Model, the Real-Time Feature Computation, as well as the Radar Screen, which is also used to log additional aircraft parameters for subsequent data evaluations, were implemented on top of the existing simulator environment. The PFDs were modified, as defined in section 3.2.4, to visualize the results of the Unstable Approach Prediction Model. All computers communicate over the User Datagram Protocol in a closed network.

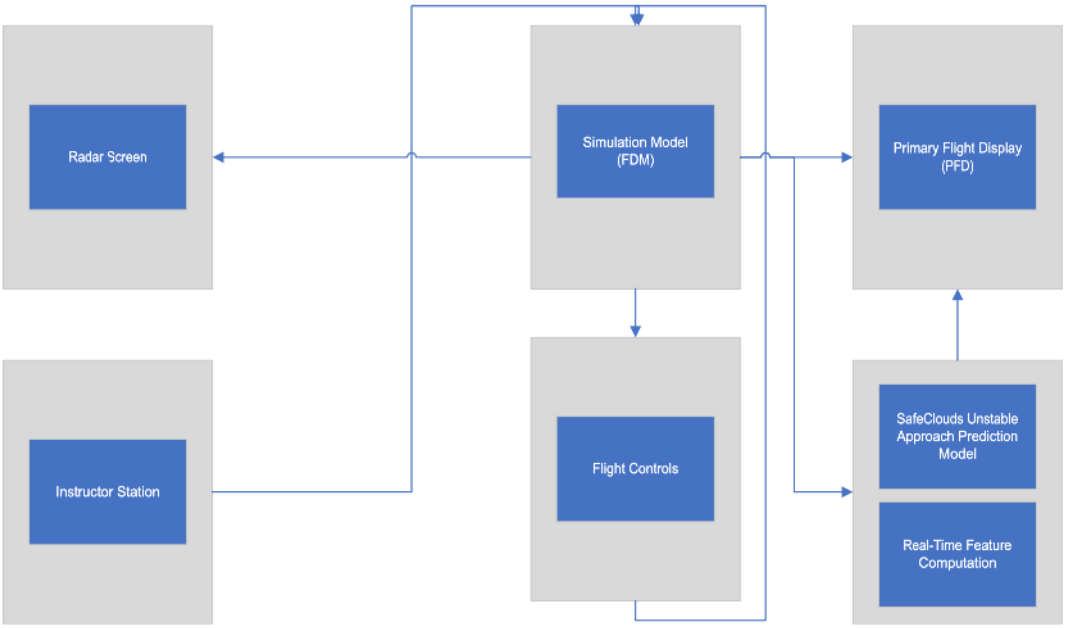


Figure 2: Scheme of the PCs in the Simulator Environment

The following subsections describe the relevant parts of the implementation in more detail.

3.2.1 Flight Dynamics Model

The flight dynamics model of the simulator is implemented in Simulink, based on a Dornier Do-728 jet. The model is structured in several subsystems, covering

- Actuators,
- Aerodynamics,
- Engines,
- Environment,
- Landing Gear,
- Weight and Balance, and
- Sensors.

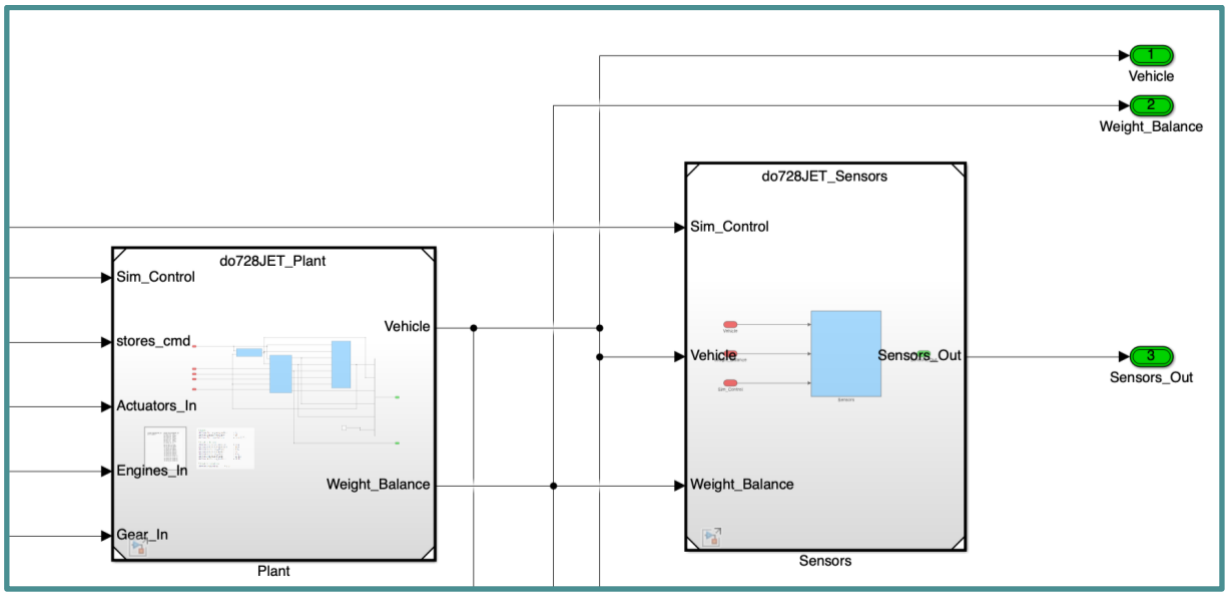


Figure 3: Simulink High-Level Model of Do-728 Plant and Sensors

The implementation in Simulink, illustrated by Figure 3, allows modifications and extensions to the existing simulation model. For this study, this is particularly important as it allows a simple implementation of the real-time feature computation model on top of the sensor model, illustrated in Figure 4. This module connects the machine learning model, summarized in subsection 3.2.1, to the simulator.

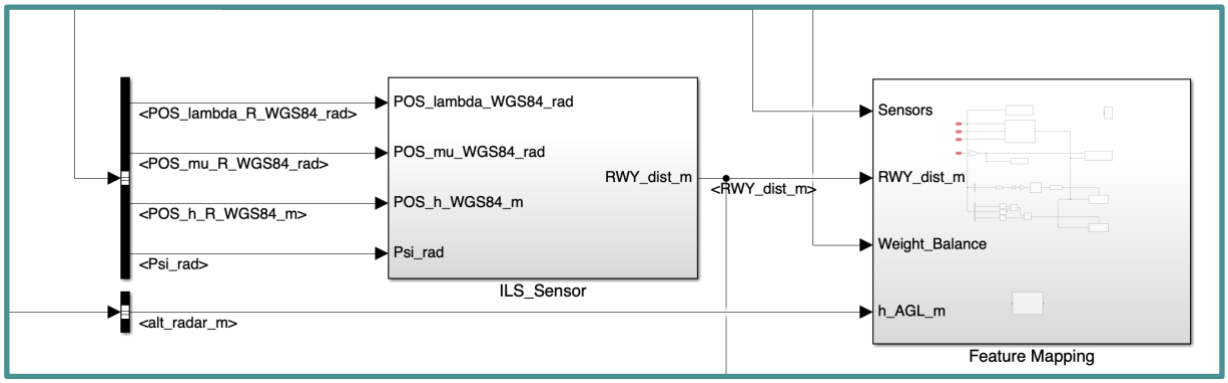


Figure 4: Simulink High-level Model of Real-Time Feature Mapping

3.2.2 Unstable Approach Prediction Model

The core of this digital assistant is a supervised, binary classification, machine-learning model that predicts unstable approaches, described in detail in D3.2 (SafeTEAM)- Section 1.2 and in (Martinez, et al., 2019).
In the following, only the relevant information for this deliverable is summarized. The focus of the summary is the prediction accuracies (recall, precision, and AUC) and the features that need to be provided to the model for a prediction.

From the initial testing of the model, based on a training/testing data split, the model achieved the accuracies, summarized in Table 1. The precision for a prediction of an unstable approach is 0.85 with a recall of 0.53, which means that if an unstable approach is predicted, the prediction is correct in 85% of the cases but only 53% of unstable approaches are predicted as such.

Table 1: Prediction Accuracies from the model testing

Class	Precision	Recall	Specificity	F1-score
Not UA	0.97	1.00	0.53	0.98
UA	0.85	0.53	0.99	0.65
Avg / Total	0.97	0.97	0.56	0.9
AUC (ROC)		0.96		
AUC (PR)		0.77		

Table 2 lists the features the model takes as input to predict the stability of the approach. Seven feature groups separate the features, providing some reasoning behind the definition of the features. These are relevant as the simulator model must be enhanced by an interface to provide this information for the implementation of the prediction model in the simulator.

Table 2: Overview of the features of the prediction model

Feature Group	Features
handling quality	pitch_rad_var, roll_rad_var, heading_rad_var, aoa_rad_var, p_radDs_var, q_radDs_var, r_radDs_var
aircraft energy	airspeed_mDs, energy_level, gndspeed_mDs, hbaro_m, hdot_mDs, mass_kg, rheight_m
adverse weather	pstatic_NDm2, wind_dir_rad, wind_spd_mDs, wind_dir_rad_var, wind_spd_mDs_var, METAR (static)
configuration	flaps_rad
crew coordination	pilot_flying (includes autopilot status)
pilot awareness	distance_m, flight_time_s, utc_time_s, number_of_holdings

To implement the prediction model, these features need to be computed in real-time, based on the aircraft model of the simulator. This model and the real-time feature computation is covered in the next subsection.
From the initial testing of the model, based on a training/testing data split, the model achieved the following accuracy.

3.2.3 Further Analysis of the Unstable Approach Prediction Model

The Unstable Approach Prediction Model was trained on real-world Flight Data Monitoring data of Airbus A320 family aircraft. The flight dynamics model of the research simulator is a Dornier Do-728 aircraft that is similar to the A320 family in the sense that it is also a two-jet engine aircraft. However, the Dornier Do-728 is shorter and lighter than the A320 family. Thus, to ensure compatibility of the simulator and the Unstable Approach Prediction Model, we performed some further analysis during the integration phase.

3.2.3.1 Simulation-Based Analysis

During the integration phase of the Unstable Approach Prediction Model in the simulator, we performed a simulation-based approach to check the compatibility of the simulator's flight dynamics model with the Unstable Approach Prediction Model. This test uses Flight Data

Monitoring data from an airline that did not contribute to the training data set of the Unstable Approach Prediction Model. The autopilot control algorithms for the simulator aircraft were modified to follow the trajectories from 88 unstable approaches found in the Flight Data Monitoring data. With this approach, we ensured that the approaches are realistic but also respect the dynamics of the simulator's flight dynamics model. Figure 5 illustrates the results of this model-based integration test. The x-axis illustrates the reason the Flight Data Monitoring Analysis labelled an approach as unstable. The blue bar shows the number of flights per category that are not predicted by the Unstable Approach Prediction Model. The red bar shows the number of approaches per category that are predicted as unstable approaches. (Uzun, 2024)

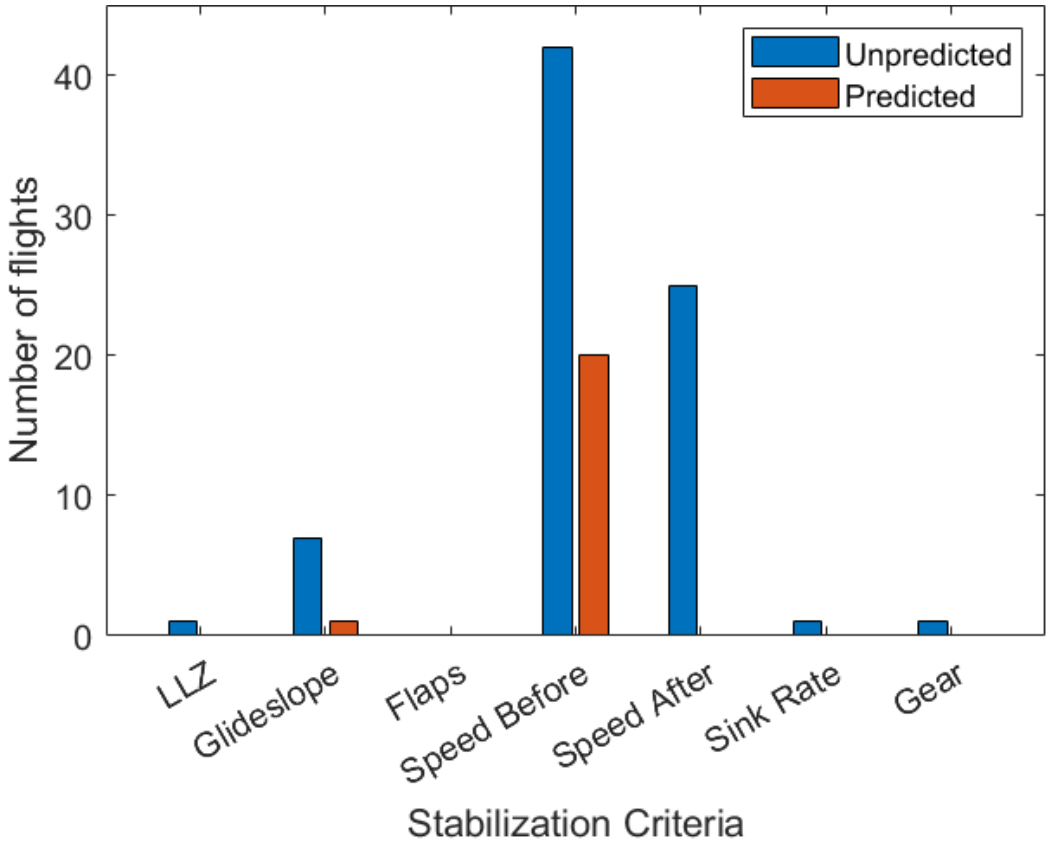


Figure 5: Results of an intermediate integration test of the Unstable Approach Prediction Model with the research simulator's flight dynamics model. The figure shows the ratio of predicted and unpredicted unstable approaches in flight data of an airline that is re-simulated with the flight dynamics model of the simulator. (Uzun, 2024)

Based on this analysis, we see that the Unstable Approach Prediction Model, fed with data generated by the simulator's flight dynamics model, can predict unstable approaches. The recall is lower than that tested in the Unstable Approach Prediction Model's test, summarized in Table 1. It is difficult to judge if this drop in recall stems from the change in airline, airport or the flight dynamics model itself, but the test nevertheless shows that the model is compatible to the simulator in the sense that it still predicts 21 as unstable approaches.

3.2.3.2 Data Driven Analysis

Additionally, we performed a data-driven test, based on recorded data from the simulator, flown by pilots. Modelling of the landing comes from two sources. Firstly, the existing ML model (Martinez, et al., 2019), trained on data gathered and processed during SafeClouds.eu (SafeClouds Consortium, kein Datum). Secondly, the data from ten landings flown in the

simulator. Thereon, we use traditional statistical methods to identify causes of unstable landings and their causes using so-called 'shallow learning'. We use the term shallow to indicate learning from data with transparent models, e.g., Principal Component Analysis and time series with (some) domain-specific knowledge.

To differentiate the work in this section from the rest of the deliverable, we use an entirely data-driven approach. From the data received, there are 29 parameters. One task is to identify which are the most relevant when detecting an unstable landing. Figure 6 shows the landing data from an agnostic data perspective, and how the parameters are correlated. For each parameter, their correlation is indicated by a value in the range of [-1.0, 1.0] where –1.0 indicates strongly negatively correlated (as one increases, the other decreases), whereas 1.0 means that as one value increases, so does the other.

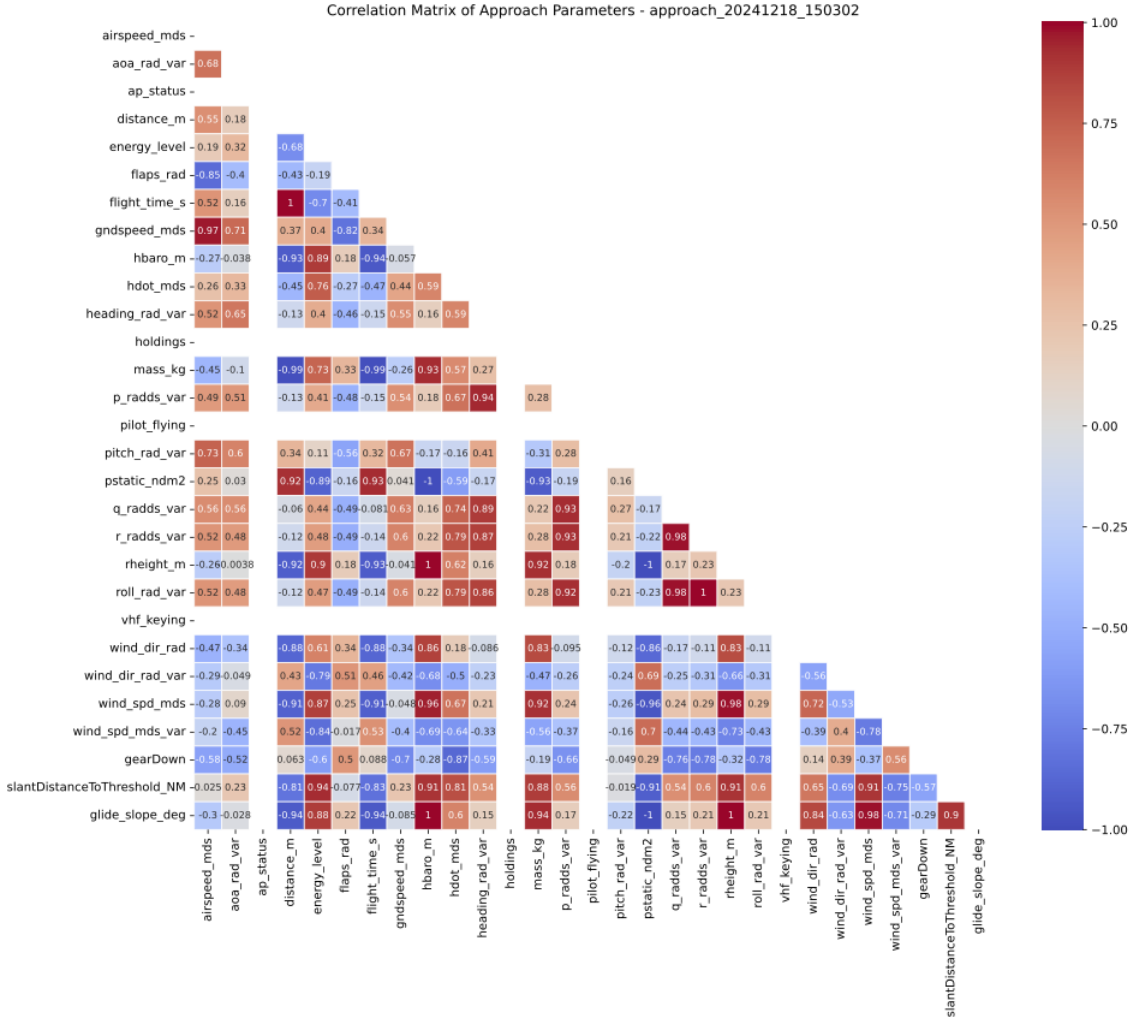


Figure 6: Correlation of the parameters from one approach flown in the simulator.

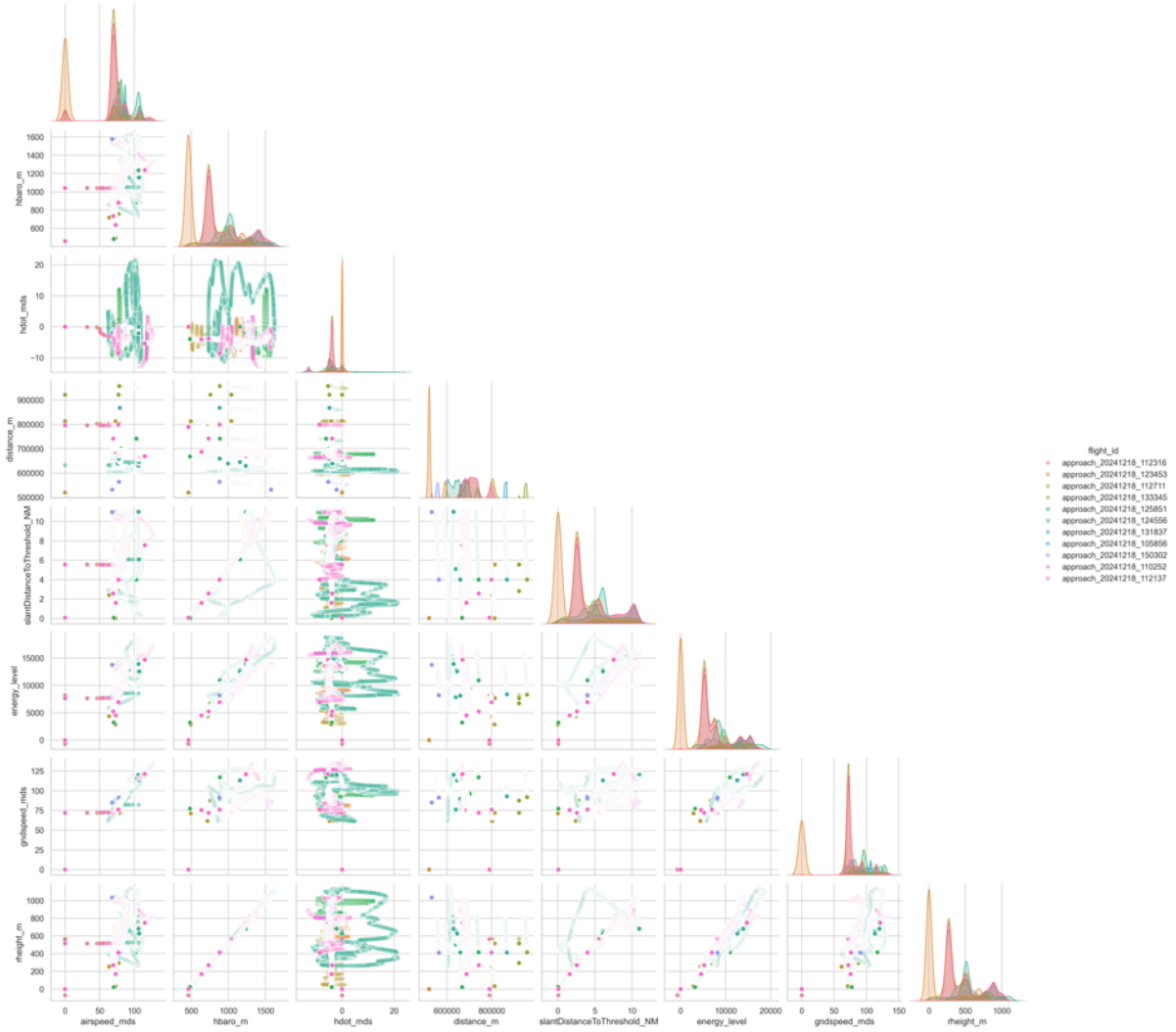


Figure 7: Parameter relationships from all approaches considered in this test.

Looking at all landings graphically, Figure 7 shows the relationship of seven parameters visually. One point of this data-driven, agnostic approach to data analysis is to identify which parameter may or may not be needed. Figure 8 shows which parameters have the largest variation with respect to the planes' energy. To keep this deliverable to the main results of this work package, we do not produce a full set of data-driven results and the background thereon. Rather, this topic will be discussed in a future paper.

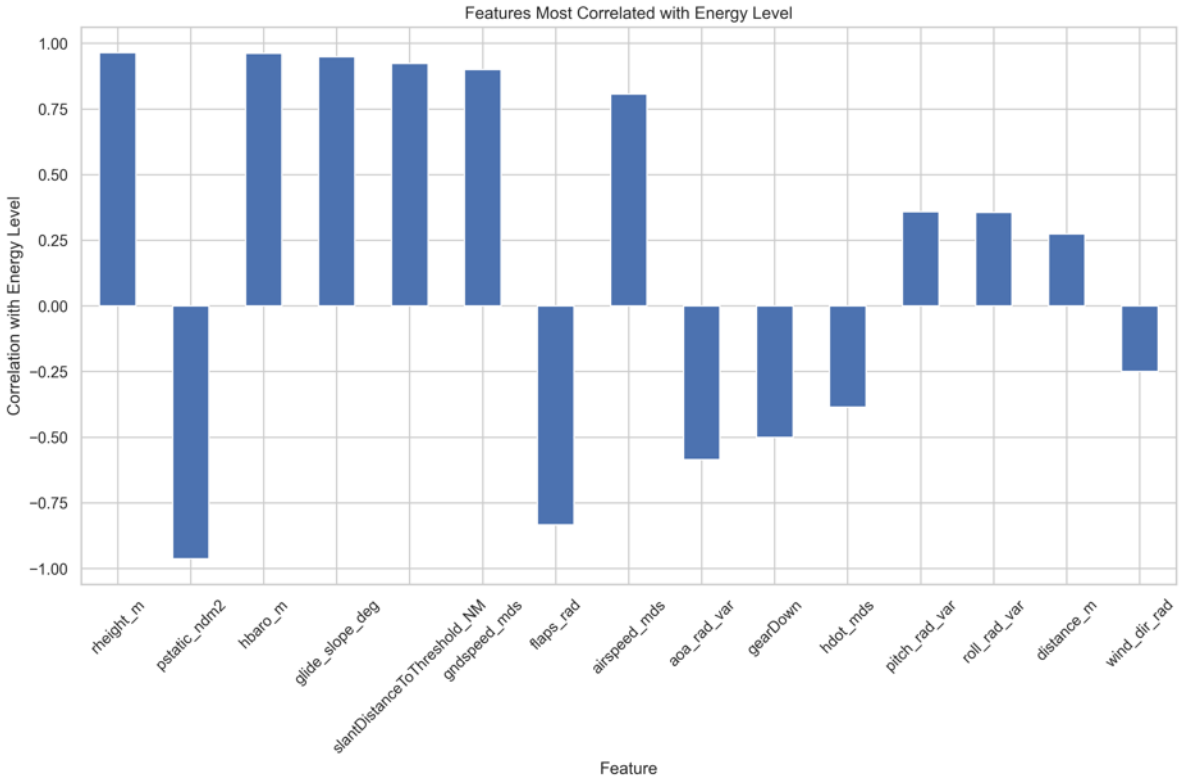


Figure 8: Example of the correlations between energy and the other parameters

3.2.3.3 Recalibration of the Prediction Threshold

Based on intermediate integration test results obtained during the implementation period, and subsequent tests of the SADA in the simulator with a pilot working on the use case team, we modified two parameters of the Unstable Approach Prediction Model, as described in section 3.2.2. First, we added an offset of +8 kts to the airspeed signal to take into account the difference in approach speed between the Airbus A320, from which the training data for the Unstable Approach Prediction Model originates and the Dornier 728, which is the aircraft modelled in the research simulator and has a lower approach speed. Secondly, to counteract the lowered recall value found in section 3.2.3.1, the threshold below which the model outcome is considered as a prediction for an unstable approach is set to 0.95.

3.2.4 Cockpit

The important part of the simulator's cockpit, for this study, is the completely modifiable PFD. Figure 9 is a photograph of the simulator's cockpit, showing the captain's PFD, in this case transparent with a green overlay to guide a 3D approach (which is not needed for this project but illustrates the design flexibility). Additionally, the Mode Control Panel for controlling autopilot modes, thrust levers, and gear handle can be seen.

Figure 9: Photo of the simulator's cockpit

Based on the design discussions and feedback from user workshops, performed within the scope of SafeTEAM's Deliverable 3.2 (SafeTEAM), the modifications of the PFD for the SADA are oriented along the design common to other caution, advisory, and warning messages like the smart runway and smart landing system.

To indicate that the prediction model, described in subsection 3.2.1, predicts an unstable approach, the PFD shows a yellow STABILIZE message, as shown in Figure 10. The text can be triggered by User Datagram Protocol signal from the prediction model to the PFD software.

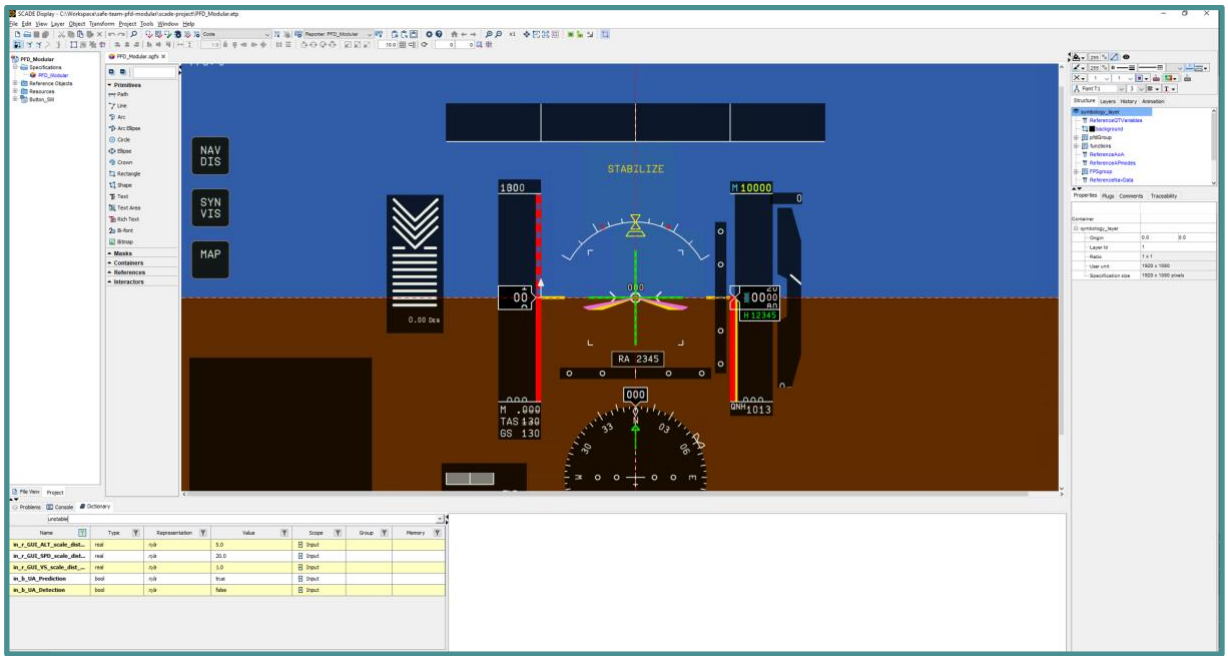


Figure 10: Adding the Stabilize Caution Message to the PFD, in Scade, the Software used to Implement the Primary Flight Displays of the Research Simulator

3.2.5 Radar Screen

For this study, we expanded the simulator with a radar screen simulation. The radar screen enables the simulation controller to provide repeatable Air Traffic Control (ATC) instructions to pilots.

Figure 11 illustrates the radar screen simulation used for SafeTEAM. The white square marker indicates the position of the aircraft, the blue text states the aircraft's callsign, and the white text shows the altitude and airspeed. The thick white lines at the intersection of two cones indicate the runways at Munich Airport. The thinner dashed lines indicate the extended runway center lines, where each dash has a length of one nautical mile. The dotted line indicates the tower control zone of Munich Airport. The cones are added to the radar screen for the operator's orientation for specific ATC requests during the simulation exercises.

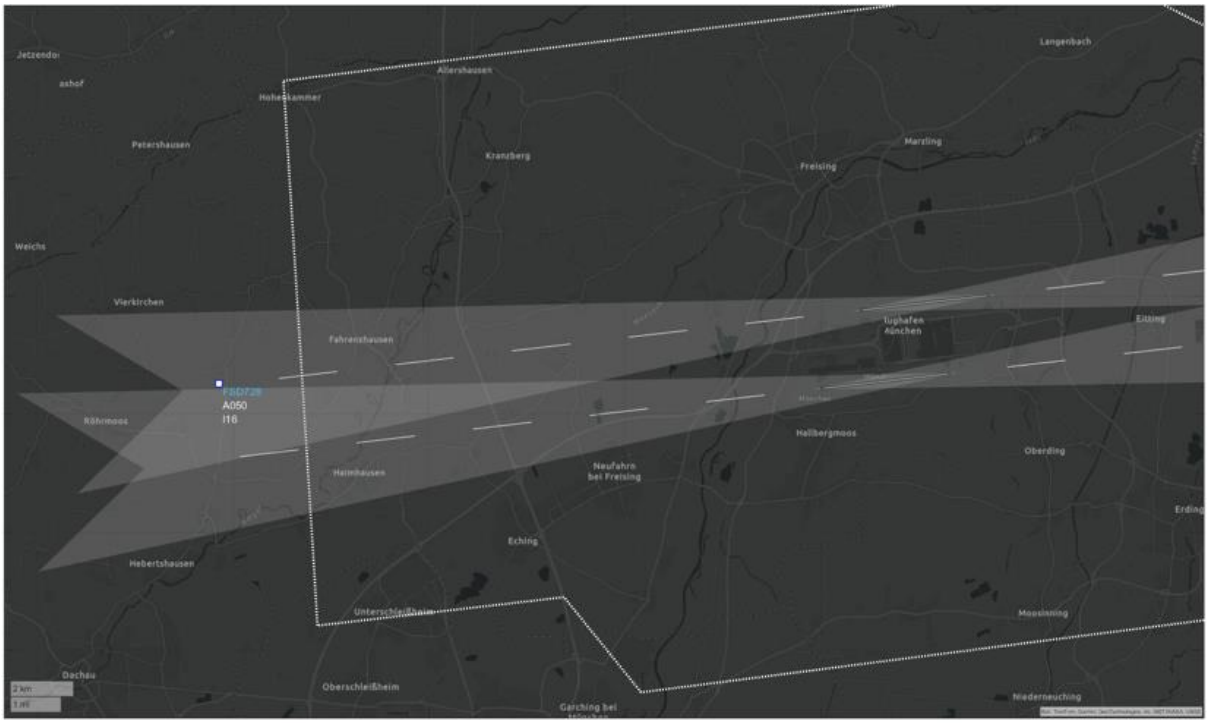


Figure 11: SafeTEAM Radar screen visualization as used for the simulation exercises. The radar simulation visualizes the aircraft position, altitude, and speed to the simulation operator in real-time. Additionally, for orientation of the operator, the cones provide visual clues when requesting Air Traffic Control instructions to the pilots during the simulation to ensure comparability of the scenarios between varying pilots. Basemap attribution: "Esri, TomTom, Garmin, Foursquare, GeoTechnologies, Inc, METI/NASA, USGS"

4 Research Question Definitions

The overall objective of the SADA is twofold. First, the assistant aims to help mitigate preventable unstable approaches. We define a preventable unstable approach as an approach that, without intervention, would violate stabilization criteria after passing the stabilization gate, but where timely pilot action (prompted by SADA) could prevent instability. Second, for approaches that are unavoidably unstable, the assistant is intended to support go-around decision compliance by a timely prediction at 4NM from the runway threshold, followed by a clear indication once stabilization criteria are violated.

Both objectives are centered on improving pilot situation awareness (SA) and decision-making during approach phases that are often ambiguous until the stabilization gate or just before. The design rationale and operational concept of the SADA system are summarized in Section 3 and detailed in Deliverable D3.2 (SafeTEAM).

Given the early-stage maturity of the SADA implementation and the constraints of this simulation-based study, it was not feasible to design a large-scale test campaign to directly evaluate these two high-level objectives. Specifically:

- A sufficiently large number of pilots and approaches would be required to generate statistically robust comparisons between assisted and unassisted conditions.
- The current implementation of SADA reflects a research-stage prototype consistent with the D3.2 specification. Further iterations and refinement are needed before large-scale validation is practical.

As such, we decomposed the overarching aims into more focused, testable research questions that address prerequisite conditions for the assistant's effectiveness. We examine whether the SADA system:

- Enhances pilot situation awareness of factors relevant to approach stability,
- Avoids increasing mental workload, and
- Is trusted and perceived as usable by its end users.

These research questions are addressed using both qualitative and quantitative data. While the small sample size limits the statistical power of quantitative comparisons, we nonetheless analyze these data to provide tentative support and nuance for findings from the analysis of post-session interview feedback.

The following table summarizes the research questions, hypotheses, as well as the methods used to test them.

Table 3: Research Questions and Hypotheses Overview

ID	Research Question	Hypotheses (if applicable)	Related Data Collection
1	What is the relationship between pilots' perception of approach stability and the SADA's perception of approach stability?	H0: no correlation H1: positive correlation H2: negative correlation	• Post-scenario questionnaire: stability self-assessment (Section 5.4.3.1) • SADA stability assessment (Section 5.4.1)
2	How does the SADA affect pilots' situational awareness during final approach (10 NM to decision height)?	H0: no correlation H1: positive correlation H2: negative correlation	• Post-scenario questionnaire: SA3 rating (Section 5.4.3.1) • Post-session interview: SA themes
3	How does the SADA affect pilots' mental workload during final approach (10 NM to decision height)?	H0: no correlation H1: positive correlation H2: negative correlation	• Post-scenario questionnaire: NASA-TLX (Section 5.4.3.1) • Post-session interview: workload themes
4	What is the effect of the SADA on go-around compliance? (as addressed in (Flight Safety Foundation, 2000))	Explorative	• Post-session interview: decision-making themes • Simulator flight data (Section 5.4.1) • SADA assessments (Section 5.4.1)
5	What is the perceived usability of the SADA system?	Explorative	• Post-session questionnaire: System Usability Scale (SUS) (Section 5.4.3.2) • Post-session interview: usability themes

5 Methodology

This section summarizes the design of the experiment. The exercise of this use case is defined for a number of five participants. The relatively low number of participants needs to be taken into account for the evaluation strategy generally and specifically when selecting metrics. Also, the selected metrics must be considered in conjunction with the simulation scenarios, as both must fit together. Therefore, the experiment design process is an iterative team effort, performed by the Human Factors experts responsible for SafeTEAM Task 2 together with the Use Case Experts responsible for SafeTEAM Tasks 3 and 4. Results in Deliverable 2.2 are built, in part, on the learnings of applying the SafeTEAM framework to this use case. The methodology description that follows in this deliverable is the result of applying the SafeTEAM framework to the SADA use case.

The following section outlines the overall methodology of the planned simulator exercise, considering the evaluation strategy, metric selection, pilot preparation material and consent forms, and the definition of the simulation scenarios.
Figure 12 illustrates the idea of designing the experiment. The upper row illustrates **independent variables**, which can be manipulated by researchers to configure experiment conditions. A specific combination of these independent variables (e.g.: approach procedure, visibility, wind, or ATC commands) defines a simulation scenario. The scenarios defined for this experiment are specified in detail in subsection 5.1. Through manipulations of **independent variables across conditions**, we expect changes in **dependent variables** like the usability of the assistant, situational awareness of pilots, or stability of the approaches. The dependent variables, the metrics to measure, and ways to obtain the metrics are defined in subsection 5.3. To collect the data of dependent variables during and after an experiment, we use three methods:

1. Questionnaires,
2. Semi-structured interviews, and
3. Data recording in the simulator.

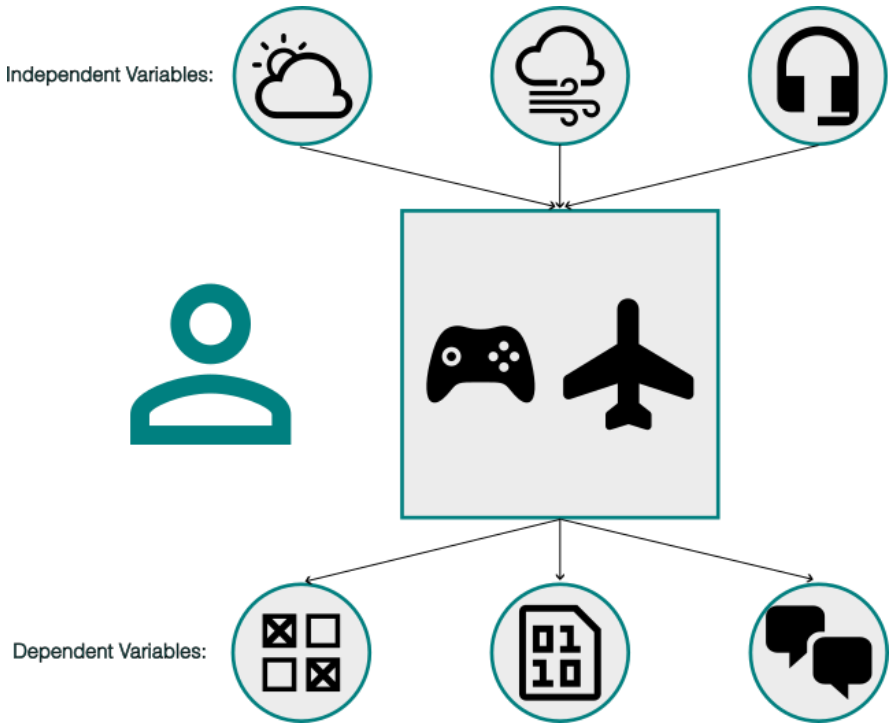


Figure 12: Schematic of the Simulation Exercise

The goal in designing the experiment is to select independent variables in such a way that the observed dependent variables provide the information to make statements on the research questions. In iterative sessions, the SafeTEAM WP2 and WP4 groups developed the simulation scenarios, as defined in section 5.1, in parallel with the data collection plan, described in subsection 5.3. These sessions focused on selecting independent variables, which define the simulation scenarios, and discussing the dependent variables and how to measure these during and after the simulation exercise.

5.1 Pilot Information Leaflet

As preparation for the pilots, we provide a pilot information leaflet that we provided to the pilots before the experiment. The leaflet describes the general idea of SafeTEAM's task 4.2 as well as a description of the simulator, basic aircraft parameters including pitch and power settings, and expected weather information during the simulation scenarios. Furthermore, it contains the relevant approach charts for the simulation exercise.

Beyond preparing for the simulation exercise, the leaflet also describes the data we aim to record during the simulation exercise and how we intend to use it.
Based thereon, the leaflet contains a consent form which the participants will need to sign, explaining that they agree with the outlined usage of the recorded data.
The complete leaflet can be found in Appendix B.

5.2 Simulation Scenarios

The challenge in designing a simulation scenario for the unstable approach use case is to create a realistic scenario for pilots in which they are uncertain about the stability of the

approach until the stabilization gate is reached. If the scenario is too simple, pilots will likely fly a stable approach without needing any decision support tools. On the other hand, if the scenario is too demanding, pilots might perceive it as unrealistic and not relevant for real-world operations. Furthermore, if the scenario creates a situation that necessarily leads to an unstable approach, pilots will have no need for a decision support tool, anticipating the instability long before, and might not even fly until the stabilization gate.
The idea is to design scenarios in which the parameters for workload in the cockpit and the complexity of the approach can be modified progressively. We define the following three **independent variables** to modify the complexity of the approach and, as a result, the workload in the cockpit:

- Wind conditions
- ATC Speed Constraints
- ATC Glide Slope Intercept

In all scenarios, pilots will perform approaches on either runway 08L or 08R in Munich, and for each difficulty level of an approach on one runway, we define a similar scenario for the other runway as well.
To have baseline scenarios to compare results against, we define **reference** and **solution** scenarios for the simulator trials. *Reference* scenarios are scenarios *without* the developed Stabilized Approach Digital Assistant (SADA). Conversely, *solution* scenarios are *with* the developed Stabilized Approach Assistant (SADA). For each solution scenario, we perform a comparable reference scenario with similar initial conditions and parameters.
Table 4 provides an overview of the six scenarios for the simulation exercise, providing a scenario ID, a short name of the scenario, and the link to the detailed definition of the scenario in the subsequent subsections. The detailed scenario descriptions contain the initial state of the aircraft, the ATC commands issued to the pilot, and the wind profile.

Table 4: Scenario Overview

Scenario ID	Scenario Name	Link to Scenario Description
1	Standard Intercept 08L	Subsection 5.2.1
2	Standard Intercept 08R	Subsection 5.2.2
3.	Base Intercept 08L	Subsection 5.2.3
4	Base Intercept 08R	Subsection 5.2.4
5	Short Intercept 08L	Subsection 5.2.5
6	Short Intercept 08R	Subsection 5.2.6

The **relevant charts** for the scenario are available at DFS' AIP:

- GPS / FMS RNAV Arrival Chart for runway 08L: https://aip.dfs.de/BasicIFR/pages/P00A8C.html
- ILS 08L: https://aip.dfs.de/BasicIFR/pages/P00A92.html
- GPS / FMS RNAV Arrival Chart for runway 08R: https://aip.dfs.de/BasicIFR/pages/P00A90.html
- ILS 08R: https://aip.dfs.de/BasicIFR/pages/P00A93.html

5.2.1 Standard Intercept 08L

This scenario is a standard approach to runway 08L, starting from the final approach fix, as defined on the Final Approach Chart for Runway 08L. The pilot can configure the aircraft

without any ATC requests and receives a timely approach and landing clearance for runway 08L. Table 5 lists the initial aircraft states for this scenario. The aircraft is already on the localizer and intercepts the glide slope from below at 5 000 ft above mean sea level.

Table 5: Initial Aircraft States for Standard Approach on 08L

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.3415	11.4964	200	5000	82

Figure 13 illustrates the wind profile for this scenario. With decreasing altitude, the wind becomes slower but builds an increasing tailwind component.

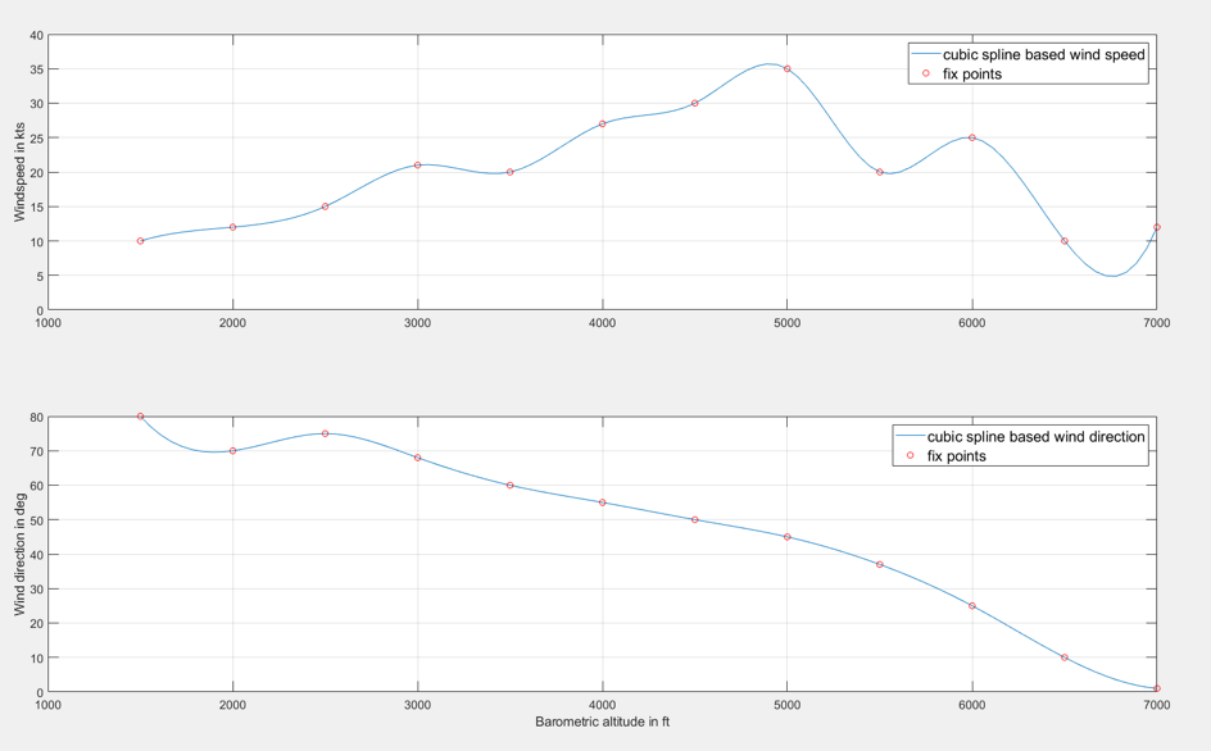


Figure 13: Wind profile for the standard approach on runways 08L and 08R

5.2.2 Standard Intercept 08R

This scenario is a standard approach to runway 08R, starting from the final approach fix, as defined on the Final Approach Chart for Runway 08R. The pilot can configure the aircraft without any ATC requests and receive a timely approach and landing clearance for runway 08R.
Table 6 lists the initial aircraft states for this scenario. The aircraft is already on the localizer and intercepts the glide slope from below at 5 000 ft above mean sea level.

Table 6: Initial Aircraft States for Standard Approach on 08R

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.3195	11.4815	200	5000	82

The wind situation is similar to the one for the standard approach on runway 08L, described in section 5.2.1, Figure 13.

5.2.3 Base Intercept 08L

This scenario shall recreate a situation in which the air traffic controller requests the pilot to maintain speed until 4NM from the threshold of runway 08L. The initial aircraft state is provided in Table 7.

Table 7: Initial Aircraft States for Base Intercept on 08L

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.4083	11.3917	220	5000	170

The **wind profile** for this scenario is illustrated in the following Figure 14.

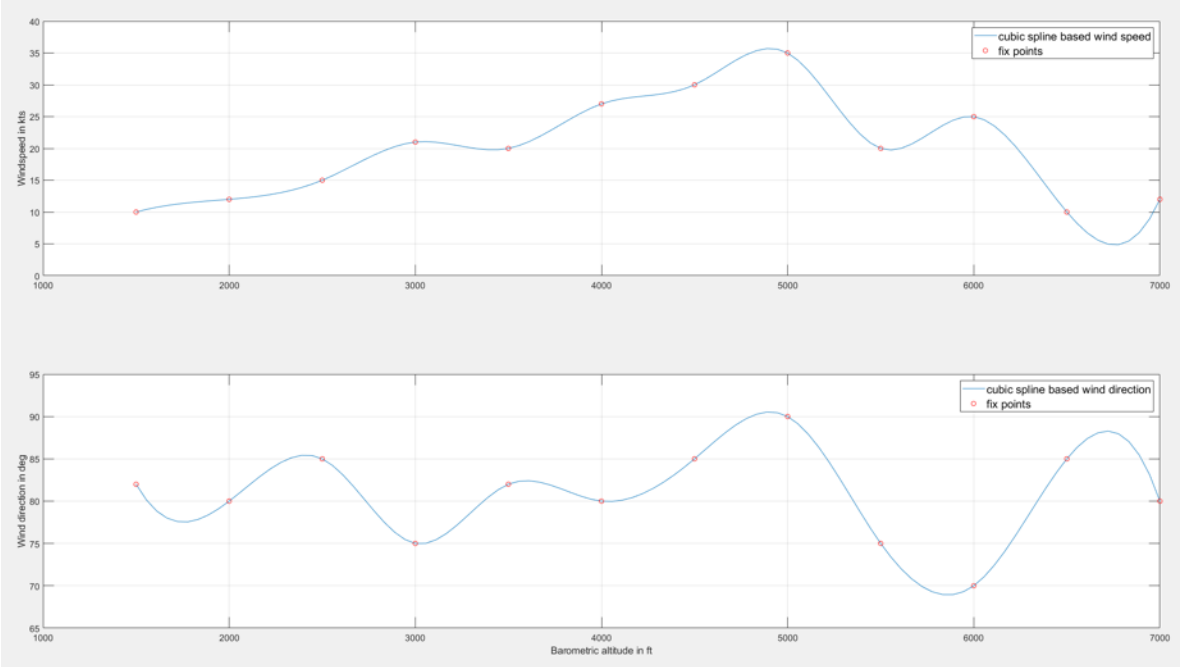


Figure 14: Wind profile for the base intercept scenario

The **ATC instructions** are:

1. Turn left, heading 120.
2. Cleared for approach.
3. Maintain 150 to 4NM.
4. Cleared for landing.

5.2.4 Base Intercept 08R

This scenario shall recreate a situation in which the air traffic controller requests the pilot to maintain speed until 4NM from the threshold of runway 08R, combined with a tailwind component. The initial aircraft state is defined in Table 8.

Table 8: Initial Aircraft States for Base Intercept on 08R

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.2417	11.3917	220	5000	350

The **wind profile** for this scenario is similar to the wind profile in scenario 3 and is illustrated Figure 14.

The **ATC instructions** are:

1. Turn right, heading 050.

2. Cleared for approach.
3. Maintain 150 to 4NM.
4. Cleared for landing.

5.2.5 Short Intercept 08L

This scenario shall recreate a situation in which the air traffic controller offers the pilot a shortcut to the final approach. On the downwind, the ATC will vector the pilot, intending to have the arrival aircraft intercept the Instrument Landing System (ILS) around 6NM from the runway threshold at an altitude of 3700 ft.

Table 9: Initial Aircraft States for Short Intercept on 08L

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.4448	11.7247	220 kts IAS	4000	262

The **wind profile** for this scenario is illustrated Figure 15.

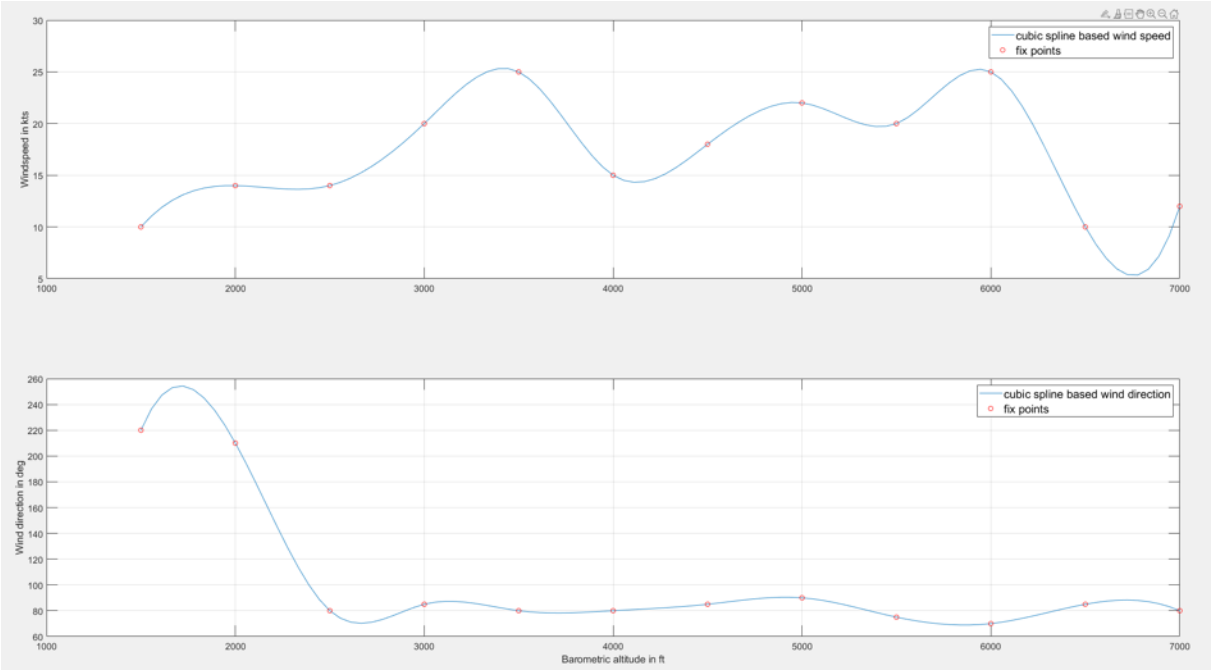


Figure 15: Wind profile for the short intercept scenario

The **ATC instructions** in this scenario are:

1. Turn left, heading 170, and descend to 3700ft.
2. Turn left, heading 120, cleared for approach.
3. Cleared to land.

5.2.6 Short Intercept 08R

This scenario is similar to the scenario described in 5.2.5, except it approaches runway 08R instead of runway 08L. The initial aircraft state is defined in the following Table 10.

Table 10: Initial Aircraft Sates for Short Intercept on 08R

Latitude [°]	Longitude [°]	Velocity [kts IAS]	Altitude [ft]	Heading [°]
48.2463	11.6565	220 kts IAS	4000	262

The **wind profile** is similar to the one of scenario 5, illustrated in Figure 15.

The **ATC instructions** are

1. Right turn, heading 350 and descent to 3700ft.
2. Right turn, heading 040, cleared to approach.
3. Cleared to land.

5.3 Scenario Ordering

The scenario ordering is determined based on a computer script that uses random number generators. Section 5.1 defines the six scenarios used for the simulation exercise of which two belong to one of three difficulty levels respectively. The script ensures that one of the similar scenarios is selected as a reference and solution scenario. Additionally, the script randomizes the order of the three selected solutions and reference scenarios.
Thus, randomization ensures that each pilot has three reference scenarios and three solution scenarios, with one scenario of each difficulty level. Additionally, the ordering of the difficulty levels is randomized for the solution as well as reference scenarios.
Furthermore, which of the two scenarios of similar difficulty is chosen as reference and solution scenario is also chosen randomly for each pilot anew.

5.4 Data Collection

Section 4 defines the research questions that this deliverable investigates. Based thereon, section 5.1 describes the scenarios pilots fly in the simulator to test the SADA, implemented as described in section 3.2. Measuring the SADA's impact on the research questions requires metrics on the Human-Machine Teaming, as well as aircraft performance indicators of the approach scenarios flown in the simulator. Especially, the metrics selection targeting Human-Machine Teaming is an evolutionary process that was performed in parallel to the development of simulation scenarios defined in section 5.1.

The selection process for the Human-Machine Teaming Metrics is based on the collection of potentially applicable metrics, included in the SafeTEAM framework, described in SafeTEAM Deliverable 2.2. These metrics aim to capture the collaboration or teamwork between Digital Assistants and the human user or operators. The complete collection of metrics is too extensive to reproduce in this deliverable; however, the metrics can be grouped by measuring the following dependent variables:

- Trust
- Workload
- Situation Awareness
- System Usability

The following subsections summarize, by grouping the data collection methods, the metrics chosen to collect data on a dependent variable.

5.4.1 Unstable Approach Prediction Model

One information type recorded in the simulator exercise are the features fed into the prediction model, computed for each approach flown, and also the resulting Unstable Approach Prediction Model's outcome. This outcome contains the output of the model as

well as the five most relevant features for the model to yield the prediction on the approach stability. An exemplary outcome, encoded as JavaScript Object Notation (JSON), can be, for example,


```
"prediction": "[0.0]",
 "probability": "[array([0.97711212, 0.02288788])]",
 "top_features": [
  ["weather_altimeter_hpa",
   1.3686779476318494],
  ["feature_4_0_nm_airspeed_mds",
   0.9550524174502503],
  ["feature_flap_full_hbaro_m",
   0.8137647769602466],
  ["feature_4_0_nm_roll_rad_var",
   0.33292498041475493],
  ["date",
   0.3299271467598154]
 ]
```


which contains key-value pairs describing the "prediction" of an unstable approach as boolean false or true variable, the "probability" as a number between zero to one which is compared to the defined threshold in section 3.2.3, to make a prediction, as well as the five most influential features provided to the Unstable Approach Prediction Model to calculate the probability outcome.

5.4.2 Simulator Data

As described in section 3.2.5, the simulator stores performance data of the aircraft, as well as pilot inputs. This allows for an analysis similar to flight data monitoring analysis that airlines do in their operations, taking into account the speed, Instrument Landing System deviations, sink rate, and aircraft configurations.
Figure 16 exemplarily illustrates the Indicated Airspeed and also the deviation from the target speed from eleven miles from the runway threshold to the runway threshold. The red box illustrates the critical region for stability analysis, indicating the region from the stabilization gate to the runway threshold in the horizontal and the parameters' stability domain in the vertical direction. In the following picture, the approach is unstable, since the target speed is not in the required range at the beginning of the stabilization gate.

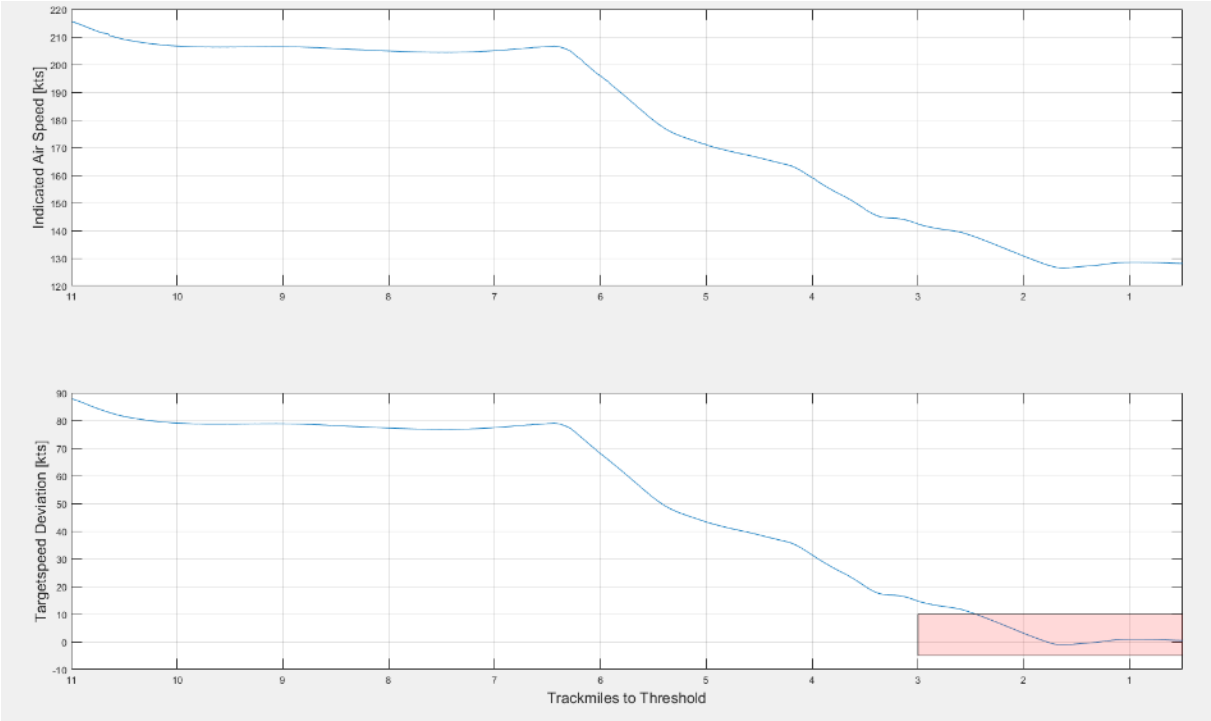


Figure 16: Approach Speed and Speed deviations during an approach

Figure 17 exemplarily illustrates the glide slope and localizer deviation of an approach from eleven nautical miles until the runway threshold. Again, the red regions illustrate the critical region to evaluate the stability of the approach. For the ILS, the deviation is measured in dots, where a deviation of one dot in each direction is acceptable, and deviations beyond one dot after the stabilization gate are considered as an unstable approach.

Figure 18 exemplarily illustrates the sink rate of an approach. The red region illustrates the region in which the parameter is considered stable. Additionally, the green line indicates the sink rate the aircraft would have if it followed the vertical guidance of the glide slope with its current ground speed.

Figure 19 exemplarily illustrates the aircraft's configuration during an approach. The gear is illustrated as a Boolean where True indicates the gear is down. The flap setting is illustrated from zero to four, where zero indicates flaps retracted and four indicates "flaps full" setting. The red regions indicate the gear and flap settings necessary to consider the approach stable.

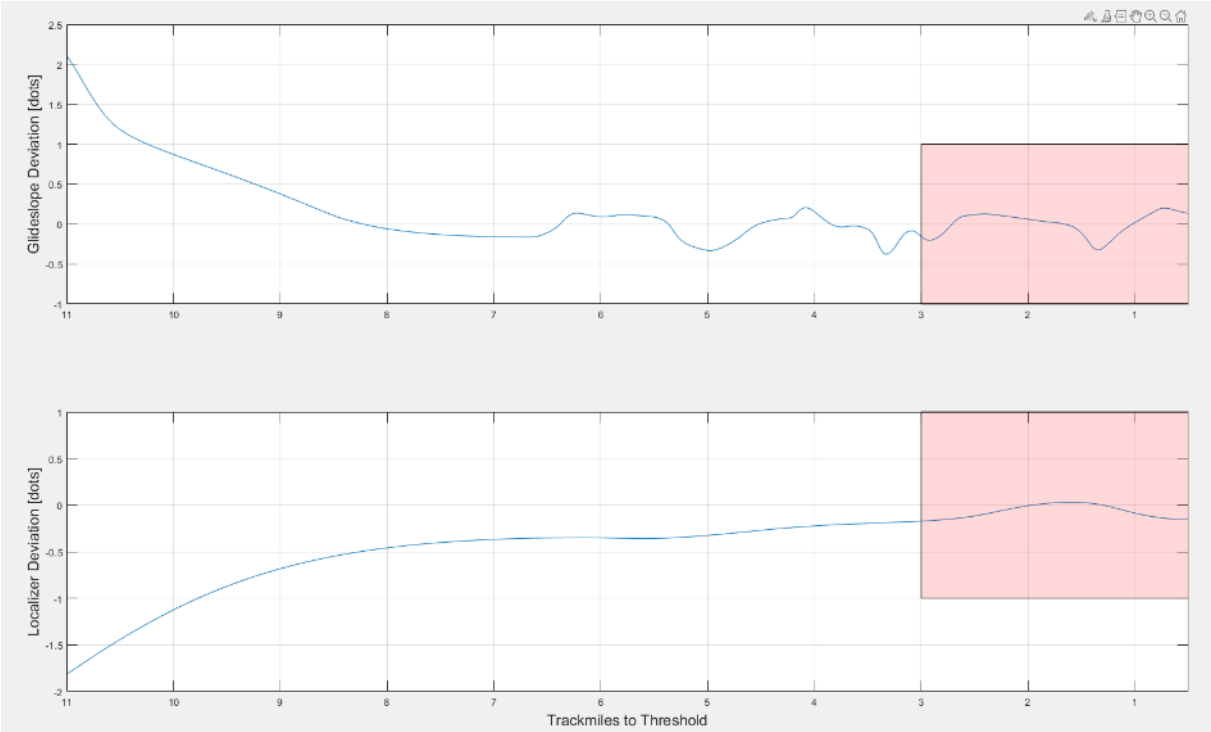


Figure 17: Localizer and Glideslope deviation during an approach

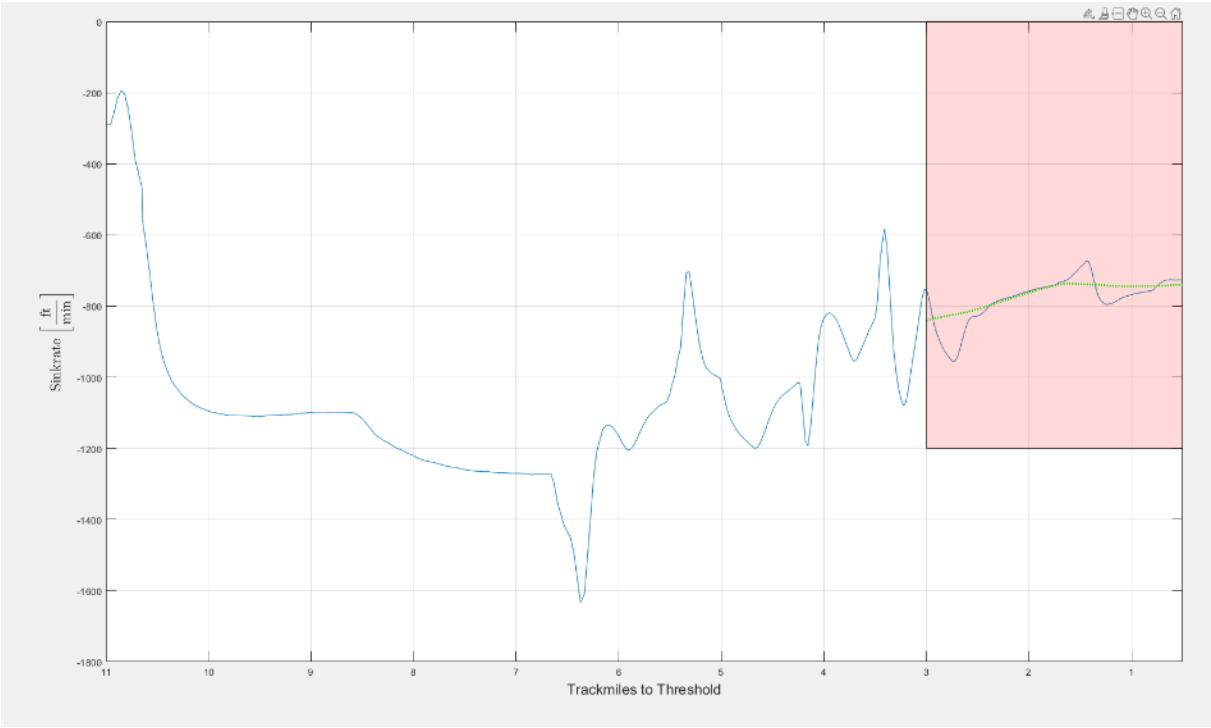


Figure 18: Sink rate during an approach

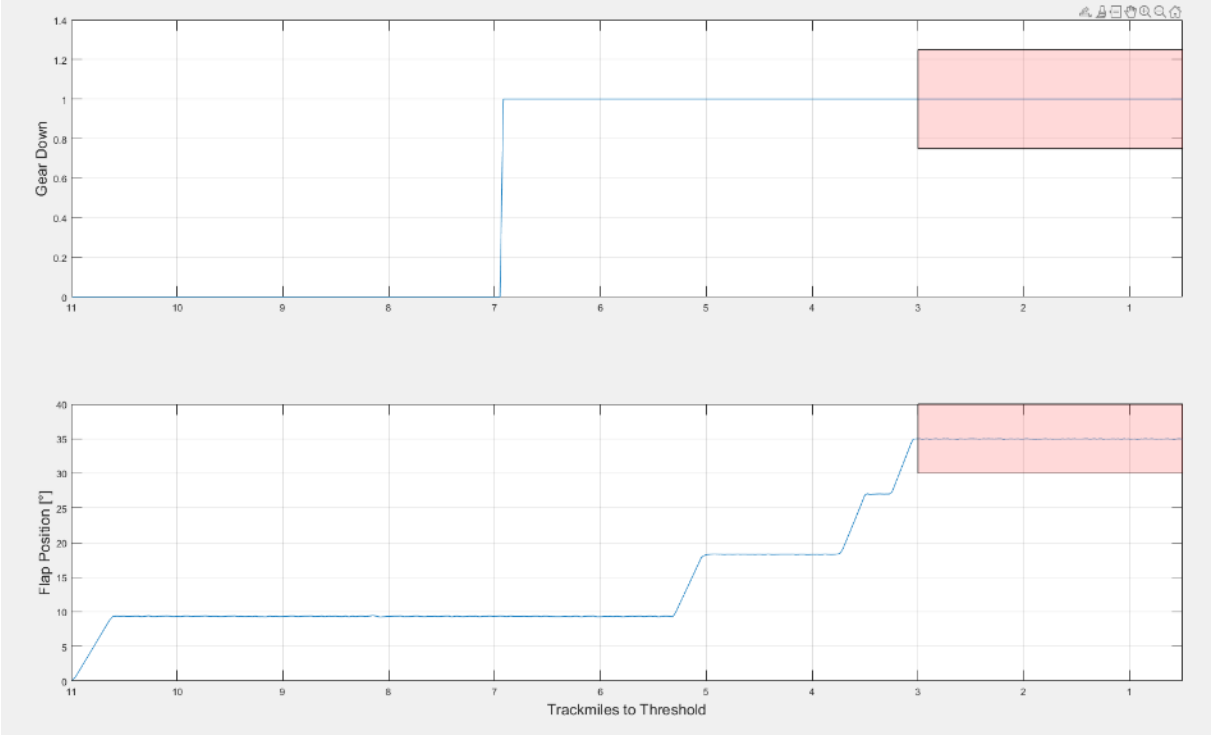


Figure 19: Aircraft configuration during an approach

5.4.3 Questionnaires

We designed two questionnaires for the SADA exercise, using the EUSurvey service (DG Digit, kein Datum). First, a post-scenario questionnaire that pilots shall file immediately once each of the six scenarios is completed. Second, a post-session questionnaire that pilots shall file after the sixth post-scenario questionnaire is completed, concluding the scenario session. The post-scenario questionnaire focuses on the dependent variables **Workload** and **Situation Awareness,** along with perceived scenario difficulty and other qualities of the scenario design itself. The collected questionnaire serves to answer the research questions, while the additional scenario "meta data" enables data validity and reliability assessments. Subsection 5.4.3.1 provides details on the post-scenario questionnaire, while Appendix A.1 provides screenshots of the complete layout and design of the questionnaire.
The post-session questionnaire targets the system usability of the SADA. Subsection 5.4.3.2 provides details on this questionnaire, while Appendix A.2 provides the complete questionnaire.

5.4.3.1 Post-Scenario

The post-scenario questionnaire is designed to assess the pilot's experience and perception of the DA after they have completed a scenario. The pilot is asked to reflect on various aspects of their experience, including the system's impact on their workload, situation awareness, and subsequent decision-making during the scenario.

Workload

To collect data on the perceived workload during the scenario, the questionnaire integrates the NASA Task Load Index (NASA-TLX) (Hart & Staveland, Hart, Staveland – Development of NASA-TLX Task Load), containing six questions that are all answered using a 0—20 Likert scale. For five questions, zero indicates "very low," while 20 indicates "very high."

- How mentally demanding was the task?
- How physically demanding was the task?
- How hurried or rushed was the pace of the task?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

For the sixth question, zero is labelled "perfect" and 20 is labelled "failure."

- How successful were you in accomplishing what you were asked to do?

Rather than deploying the standard but complex way of weighing the TLX dimensions to develop a score, our experiment uses the widely adopted Raw TLX (R-TLX) method of averaging the dimension scores to create an estimate of overall perceived workload (Hart, Hart 2006 – Nasa-Task Load Index NASA-TLX, 2006).

Situation Awareness

The questionnaire also contains three items developed by (Braarud, 2021) to target the three levels of situation awareness (SA)—perception, comprehension, and projection, following Endsley's original model of SA (Endsley, 1995) - while minimizing the construct overlap with NASA-TLX. The three Situation Awareness Three Levels (SA3) Likert items, evaluated on a 1—11 scale, are (with endpoint labels in parentheses):

- My observation of critical information (1 = "Missed important information", 11 = "Identified all needed information")
- My understanding of what was going on (1 = "Did not make sense to me", 11 = "Fully understood")
- I could look ahead and foresee what was going to happen (1 = "Could not predict", 11 = "Very accurately")

The SA3 score was calculated by averaging the response scores of the three constituent items.

Approach Stability Assessment

Given the limited sample size and the participants' limited exposure to the Stabilized Approach Assistant across the scenarios, it is difficult to measure trust in the system directly. Instead, and as a precursor for trust, the experiment aims to compare the pilots' approach stability assessments with those of the Stabilized Approach Assistant, i.e., their level of agreement.

The first item of the questionnaire, targeting situation assessment, is a **self-assessment of the approach's stability.** To gather this information, the questionnaire uses a continuous slider scale between zero ("completely unstable") and one ("completely stable") with instructions to interpret 0—0.4 as unstable, 0.4—0.6 as a grey zone, and 0.6—1 as stable. Additionally, it is possible to select reasons for instability, in case the self-assessment is an unstable approach. The predefined checkboxes included reasons like speed, glideslope deviation, sink rate, etc., and an "other" option with a text field to provide a custom response.

The second item targeting situational assessment covers the **decision support** aspect of the digital assistant. Therefore, the questionnaire investigates the effect of the digital assistant on the pilots' decisions during the approach. It contains three Likert items scored between 1—7. The three questions were (endpoint labels in parentheses):

1. To what extent did the Unstable Approach Prediction tool help you maintain a stable approach? (1 = "Not at all", 7 = "A great deal")
2. To what extent did the Unstable Approach Prediction tool influence your decision to initiate a go-around? (1 = "Not at all", 7 = "Very much")
3. Overall, how strongly did you agree or disagree with the Unstable Approach Prediction tool's approach stability assessment in this scenario? (1 = "Strongly disagree", 7 = "Strongly agree")

For each question, the questionnaire instructed participants to interpret and use a score of 4 to indicate a neutral response if they were unsure or had no opinion. They could also use a complimentary text field to explain if and why any of the questions were not applicable. These decision support questions were conditionally asked after scenarios where the Stabilized Approach Assistant was available to participants, i.e., the solution scenarios.

Scenario Meta Information

The scenario difficulty is the first item interrogated in this part of the questionnaire, using the Single Ease Question (SEQ) instrument (Sauro & Dumas), a Likert scale item asking "Overall, this task was…" with responses ranging from 1 ("very easy") to 7 ("very difficult") to assess approach scenario difficulty.

The second item in this part of the questionnaire probed the frequency of a scenario occurring in real-world operations, "How often does something comparable happen to you during work?" Participants could select one of the following options: *never*; *rarely*; *from time to time*; *frequently*; and *very often*.

5.4.3.2 Post-Session

The post-session questionnaire focuses on the usability of the SADA and employs the widely adopted System Usability Scale (SUS) (Brooke) (Lewis, 2018) to measure this. SUS consists of ten 5-point Likert items, where 1 = *strongly disagree* and 5 = *strongly agree*. Response options 2—4 are unlabeled.

A system's SUS score is calculated using the following equation:

$$SUS = 2.5\left(20 + \sum SUS01, 03, 05, 07, 09 - \sum SUS02, 04, 06, 08, 10\right)$$

This formula accounts for the alternating positive and negative tone of the items and yields a unified usability score ranging from 0 to 100. SUS scores can be interpreted using adjective-based ratings:

- 0-25: Worst imaginable
- 26-50: Poor
- 51-70: OK
- 71-85: Good
- 86-100: Excellent

Or, alternatively, using a letter-grade scale:

- A: 80-100
- B: 70-79
- C: 60-69
- D: 50-59

- F: Below 50

The ten Likert items are:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

5.4.4 Interviews

After flying the scenarios and answering the questionnaires, we performed a post-session interview with the pilots. The interview is designed as a semi-structured interview with an interview guide that is provided in Appendix A.3. After performing the first interviews, we decided to switch from a minutes-based recording to a software-based recording using Audacity (version 3.7.1 (Audacity, kein Datum)) with a local model of OpenAI's Whisper—a machine-learning model-based system (Metcalfe)—to create transcripts. This setup was run entirely offline on a laptop to ensure no sensitive data was transmitted anywhere. The transcripts are corrected by the interviewer after the interview for any potential mistakes by rehearing the audio recording and complemented with layout changes for easier analysis. Only the revised transcript is used for later analysis.
The topics covered by the semi-structured interview are:

- Communication of the Assistant and Pilot
- Coordination of the Assistant and Pilot
- Workload
- Situational Awareness
- Problem Solving and Decision Making
- System Usability
- Human Autonomy Teaming
- Trust (Reliability and Agreement)

The topics are similar to the ones covered in the questionnaires, providing a redundant method of data collection with more flexibility and room for more detailed explanations from the user.

To analyze the interview data, the transcripts were analyzed for their thematic content. A Thematic Analysis (TA) workflow was developed based on TA best practices (Braun & Clarke, Braun, Clarke 2006 – Using thematic analysis in psychology, 2006) (Braun & Clarke, Braun, Clarke 2021 – One size fits all, 2021) augmented by Large Language Model (LLM) capabilities (Paoli, 2024) (Zhang, et al., 2025). This LLM-augmented analysis procedure consisted of several steps:

1) Data Familiarization
    a) Read the transcripts.
    b) Highlight compelling extracts.

c) Take note of early impressions, tensions, contradictions (e.g., using margin comments in Microsoft Word).

2) Initial LLM-Assisted Coding
   a) Submit each transcript (without human researcher's notes) to an LLM with a prompt instructing it to develop inductive thematic codes to categorize the data[1].
   b) Cross-check codes with Braun and Clarke's six-phase reflexive TA and De Paoli's open-ended prompts approach (Braun & Clarke, Braun, Clarke 2006 – Using thematic analysis in psychology, 2006) (Paoli, 2024).
   c) Merge the output spreadsheets into a single spreadsheet (adjust code IDs).
3) LLM-Augmented Code Reconciliation
   a) This is a code comparison and synthesis phase to systematically reconcile the human annotations (reflective, interpretive, intuitive) with the LLM-generated codes (based on semantic/latent meaning).
   b) Instruct an LLM to compare and integrate the human researcher notes with the LLM-generated codes in the spreadsheet[2].
   c) Manually go through and correct/edit, e.g.:
      i) Add existing human notes from the transcripts to the "Researcher Notes" column if the LLM missed something.
      ii) Double-check and/or replace quotes for more representative or illustrative ones.
      iii) (This is partly why step 1 is so important!)
4) Theme Development with LLM Support
   a) Instruct an LLM to cluster codes into themes[3].
   b) Manually go through the themes and evaluate their validity, distinction, groundedness in the data, etc.
   c) Adjust (edit/add/remove) themes, subthemes, and code-to-theme mappings as deemed appropriate.
5) Theme Refinement
   a) (Optional) Reorganize themes and clarify boundaries.
   b) Create thematic maps (by hand and/or using LLM suggestions for connections).
6) Writing and Visualization
   a) Draft a structured summary of each theme.
   b) Select compelling quotes from the data to support themes.
   c) Refine thematic map structure (i.e., hierarchical or networked).

This analysis workflow was used to develop preliminary results in section 6.5.

5.4.5 Exclusion of Teaming Metrics

[1] The prompt in Appendix C.1 was used with OpenAI's "ChatGPT o3" on 2025-06-24.
[2] The prompt in Appendix C.2 was used with OpenAI's "ChatGPT o3" on 2025-06-24.
[3] The prompt in Appendix C.3 was used with OpenAI's "ChatGPT o3" on 2025-06-25.

The nature of the SADA's concept does not contain a two-way teaming aspect but rather a one-way collaboration from the SADA towards the pilot. Metrics focusing on the two-way teaming aspect, as discussed in detail in deliverable D2.2 of SafeTEAM, were omitted for the analysis of this deliverable.

5.5 Simulation Exercise Overview

This section outlines the timeline for the conduction of one simulation exercise, structured into pre-simulation activities, the simulation exercise itself, and the post-simulation activities in Table 11.

Table 11: Simulation Exercise Timetable

rel. Time	Activity	Details
15	Arrival and Registration	Welcome participants, confirm attendance, and complete administrative paperwork.
15	Participant Orientation	Brief participants on objectives, simulator setup, and DA functionality.
	Scenario Randomization	This part is done using a Matlab Script that randomizes the scenario order, ensuring that the learning effects of the pilots during a simulator session is randomly distributed to the scenarios and not similar throughout the campaign
15	**Block 1**: Scenario 1 (Familiarization)	Give the pilot a chance to get used to the simulator
15	**Block 2.1**: Scenario 1	Run Scenario 1, followed by post-scenario questionnaire
15	**Block 2.2**: Scenario 2	Run Scenario 2, followed by post-scenario questionnaire
15	**Block 2.3**: Scenario 3	Run Scenario 3, followed by post-scenario questionnaire
10	Break	10-minute break to reset and refresh.
15	**Block 3.1**: Scenario 4	Run Scenario 4, followed by post-scenario questionnaire
15	**Block 3.2**: Scenario 5	Run Scenario 5, followed by post-scenario questionnaire
15	**Block 3.3**: Scenario 6	Run Scenario 6, followed by post-scenario questionnaire
5	Post-Session Questionnaire	Conduct the post-session surveys: Trust in Automation, System Usability Scale (SUS), and Acceptability Scale. Administer the System Usability Scale
45	Debriefing/Interview Session	Gather qualitative feedback on the DA and participant experiences through informal discussions or observational methods, without relying on surveys.

Experiment Evaluation

6 Results

The experiments, as defined in Section 5, were performed with five pilots from three different airlines. The following subsections present the data retrieved by questionnaires, interviews, and from the simulator data.

6.1 Unstable Approach Prediction Model Results

This section provides the results of the unstable approach prediction model. Figure 20 illustrates the 24 outcomes of the unstable approach prediction model for all simulated approaches as a histogram in blue. The prediction model yielded results between 0.9423 and 0.9927. From this data, we find that the SADA provided a "Stabilize" indication twice in simulation exercises.

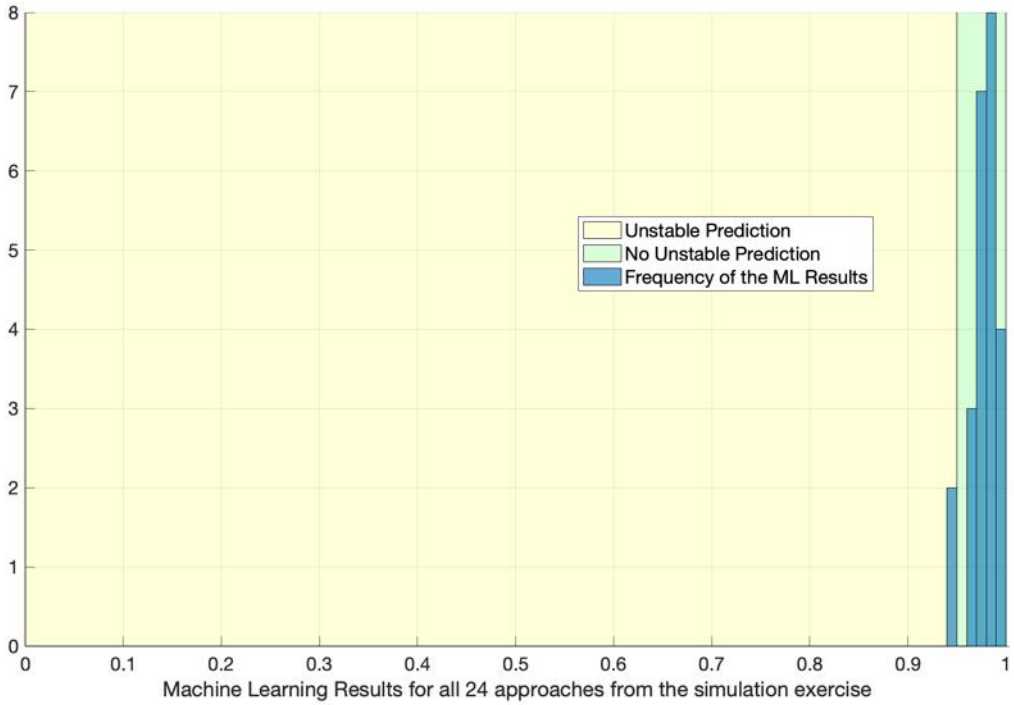


Figure 20: Unstable Approach Prediction Model Results from the Simulation Exercises

6.2 Flight Data Results

This section provides the data recorded from the simulator, as introduced in section 5.4.1. This includes all trajectories from the simulation exercises, as well as the data necessary to assess approach stability as well as go-arounds. First, the relevant parameters are visualized and described, and then summarized in Table 12

Figure 21 illustrates the two-dimensional trajectories of the approaches flown in the simulator exercises as dotted lines. The scenarios five and six, defined in sections 5.2.5 and 5.2.6, are the ones starting at the most northern and southern positions, anti-parallel to the runway direction. The scenarios three and four, defined in sections 5.2.3 and 5.2.4, are the ones starting at the north-western and south-western points, respectively. The beginning of scenarios one and two, defined in sections 5.2.1 and 5.2.2, is on the final approach fix of the northern and southern runway and is covered partly by the trajectories of scenarios three and four. The grey cones illustrate an $\pm 10°$ area around the extended runway centerlines of the runways at Munich Airport and are part of the radar screen for orientation of the simulation operator. One important observation in Figure 21 is the go-around that was flown on runway 08L.

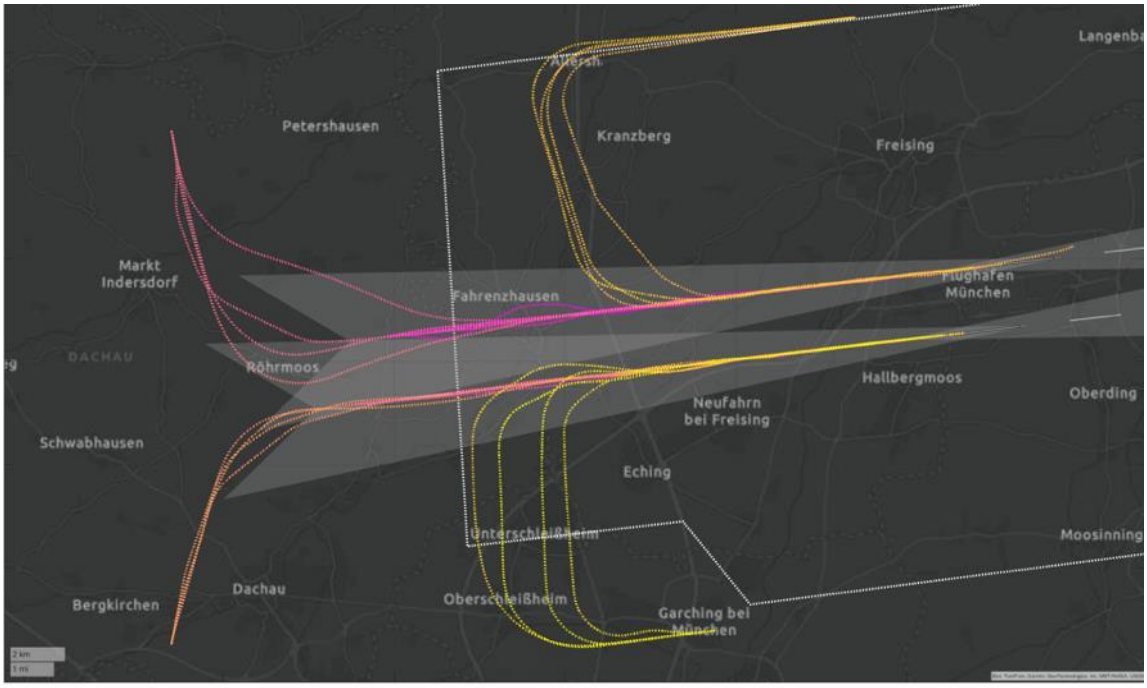


Figure 21: Trajectories of the Approaches flown in the Research Simulator

In the following, we perform a Flight Data Monitoring (FDM) analysis of the recorded data. Therefore, we check if the relevant parameters defining the stability of the approach are within the limits defined in section 2.1.4.

Figure 22 illustrates the target speed deviations on the final approach for all simulated approaches. The red box illustrates the region from $-5\ kts$ to $+10\ kts$ around the target speed from $3NM$ to $0.5\ NM$ from the runway threshold. This $3NM$ point is not precisely the stabilization gate for each approach. Since the stabilization gate is defined at $1000\ ft$ above the airfield, the stabilization gate differs for each approach. Therefore, the horizontal stretch of the box only gives a rough indication of where the stabilization gate starts, which is good enough for the subsequent evaluation of the approaches. The color coding of the target speed trajectories visualizes the stability evaluation of the unstable approach prediction model, presented in the previous section 6.1. From the analysis of the simulator data, we find seven approaches which are outside the target speed range, defined for a stabilized approach.

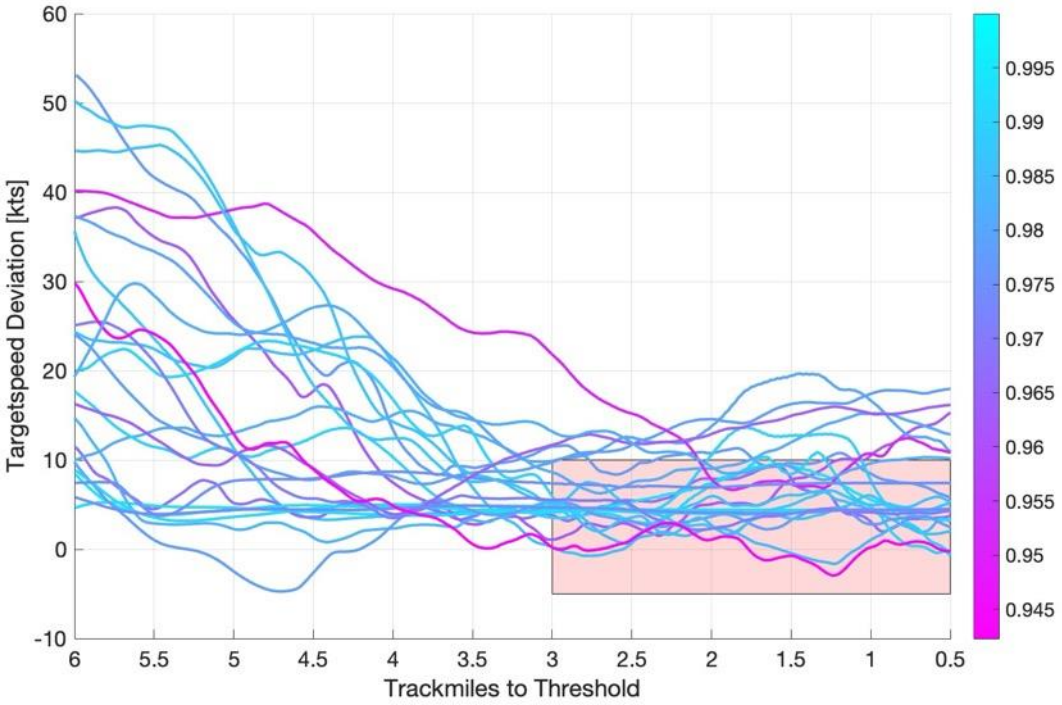


Figure 22: Target Speed Deviations on Final Approach

Figure 23 illustrates the glideslope and localizer deviations in the final approach from all simulation exercises. The red boxes indicate ILS deviations of ± 1 dots. Similar to the visualization of Figure 22, the horizontal stretch simplifies the region of the stabilization gate to $0.5NM$ to the runway threshold. Also, the color coding is the same as Figure 22. The localizer deviations are within a stable regime for all approaches except for one. The glide slope deviations, especially towards the runway threshold, are outside the domain of stability for five approaches, that do not stay within in ± 1 dot deviation of the glideslope.

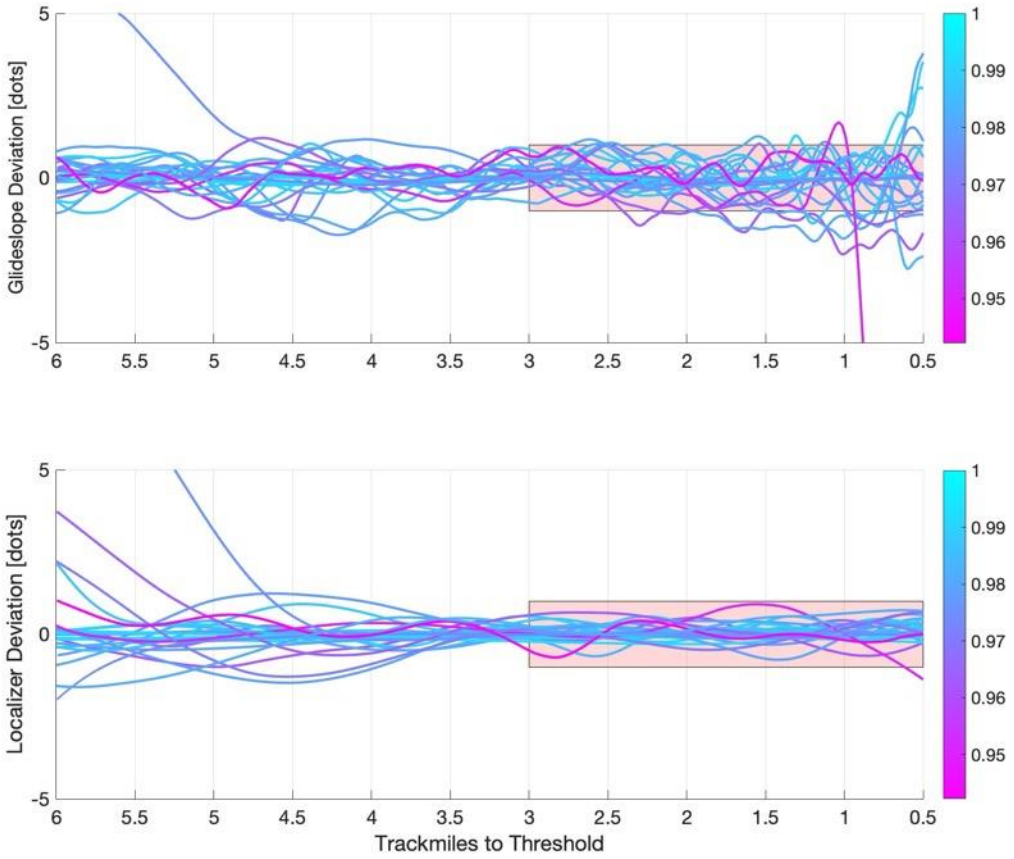


Figure 23: ILS Deviations on the Final Approach

Figure 24 illustrates the gear and flap settings during the final approach of the simulation exercises. Similar to the visualization of Figure 22, the horizontal stretch simplifies the region of the stabilization gate to $0.5NM$ to the runway threshold. Also, the color coding is the same as Figure 22. We observe that the gear configuration for all approaches was aligned with the stabilization criteria. For the flap configurations, we observe one approach, which extends the flaps after the stabilization gate. Additionally, we can observe the go-around, which retracts the flaps by one setting.

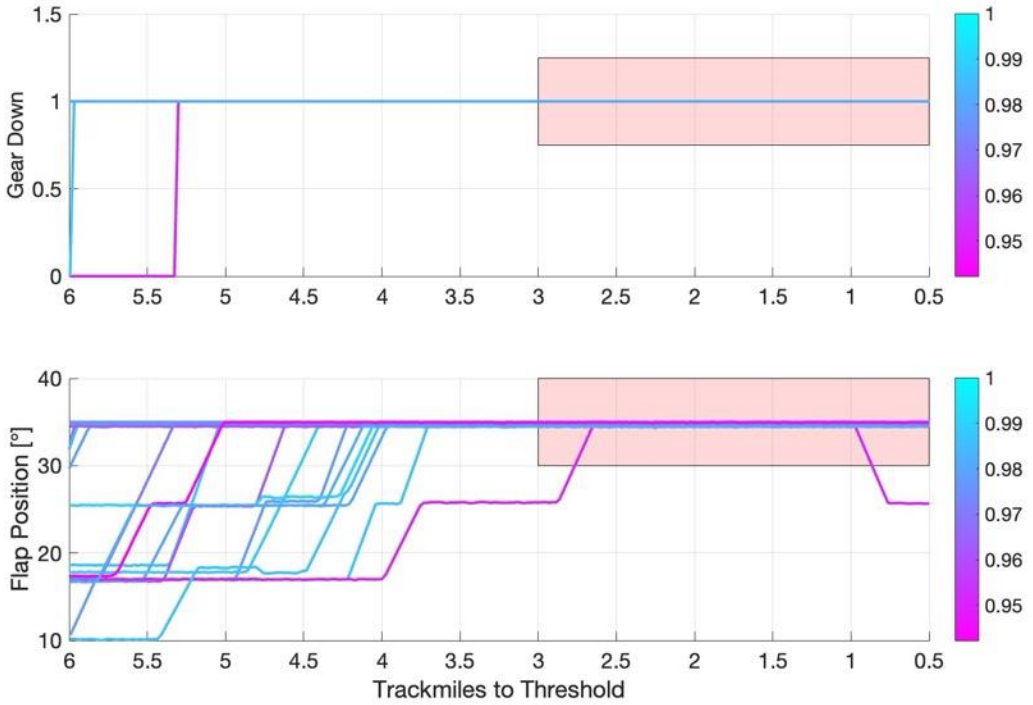


Figure 24: Aircraft Configuration on Final Approach

Table 12 summarizes the data that was recorded in the simulator and evaluated according to the stabilization criteria defined in section 2.1.4. From 24 approaches, we find 16 stable approaches and eight unstable approaches. Of the eight unstable approaches, seven were due to overspeed, combined with glideslope deviations or too late flap configuration. One unstable approach only showed a glide slope deviation without any other factors for instability. From the eight unstable approaches, one go-around was initiated.

Table 12: Flight Data Monitoring Results, Based on Simulator Data

Stable	Unstable	Speed	Glideslope	Configuration	Go-around
16	8	7	5	1	1

Finally, Figure 25 again illustrates the Unstable Approach Prediction Model Outcomes as in Figure 20, but color-coded by the FDM analysis outcomes. The red colored part of the histogram illustrates the approaches categorized as unstable. The green colored part of the histogram illustrates the approaches categorized as stable by the FDM analysis. An important observation is the green bar at around 0.95, which is a nuisance alert of the SADA. The figure also allows the computation of the SADA's precision of 50% and recall of 12.5% during the simulation exercise.

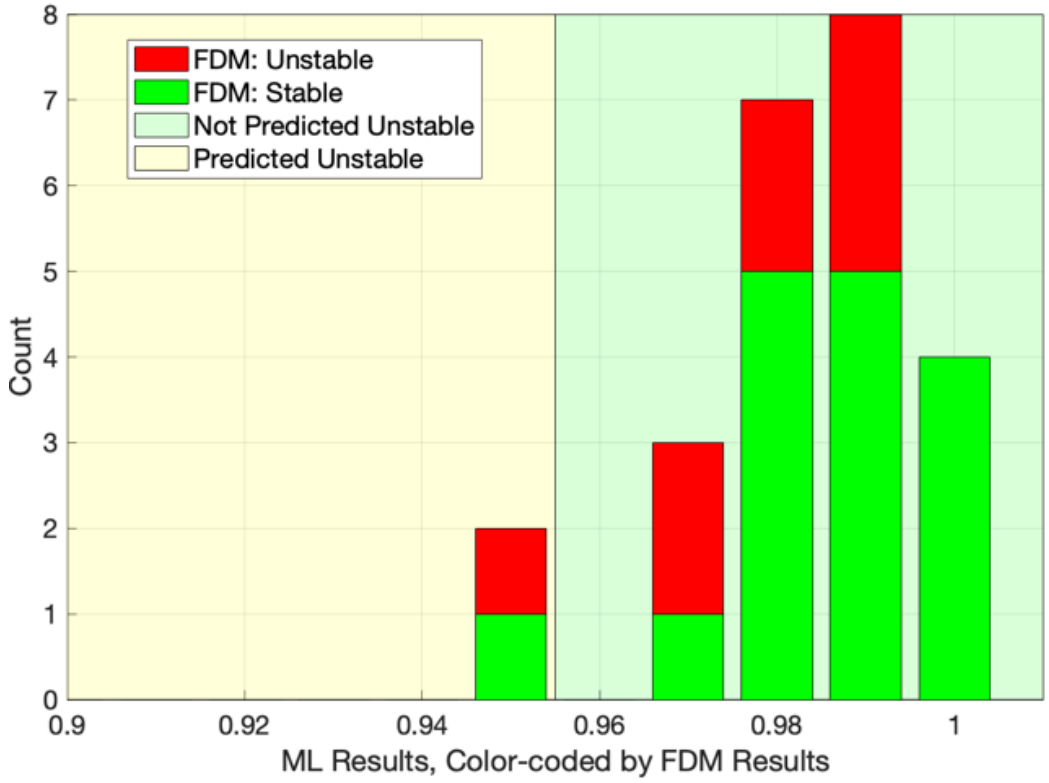


Figure 25: Unstable Approach Prediction Model Outcomes, Color-coded by FDM Outcomes

6.3 Post-Scenario Questionnaire

This section provides the results of the post-scenario questionnaire, divided by topics into subsections.

6.3.1 Stability Self-Assessment

Figure 26 illustrates the pilots' stability self-assessment of the approaches flown during the simulation exercise as a histogram plot. Additionally, the colored division indicates the scale that was defined in the questionnaire. Values from 0.0 to 0.4 indicate unstable approaches. Values from 0.6 to 1.0 indicate stable approaches. We left the area from 0.4 to 0.6 as a grey zone, as sometimes, even though approaches are categorized as unstable, the parameter that violates the stabilization criteria is out of limits only for a short period of time, for example, the speed due to a gust. In these cases, even though the approach is nominally unstable, many pilots continue the approach, making the case that the parameter violation is just short-term and under control.

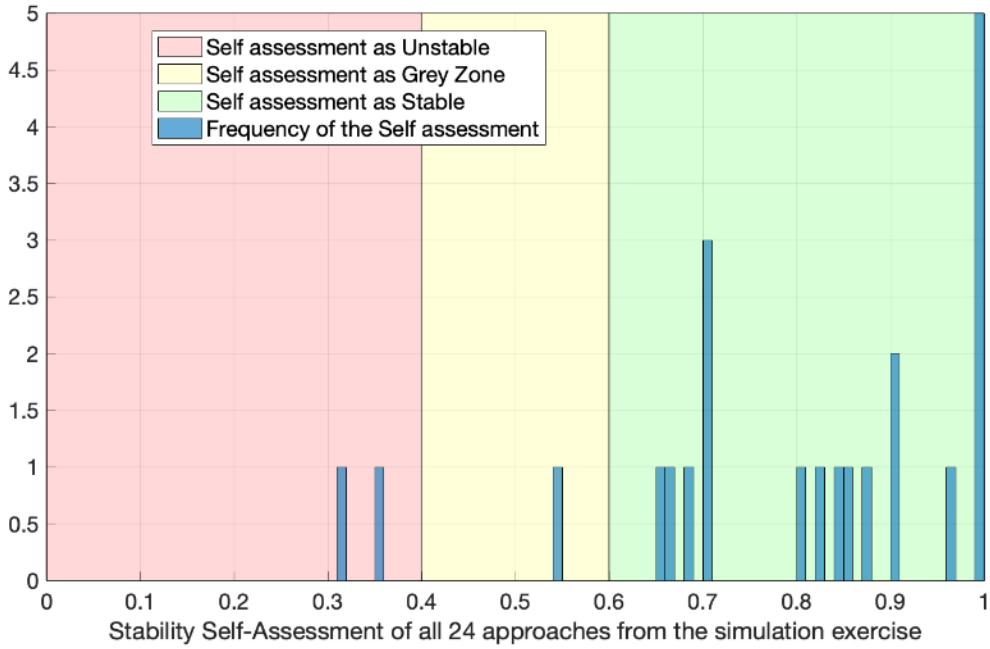


Figure 26: The pilots' stability self-assessment of the 24 approaches flown in the simulator. The x-axis is separated by colors, according to the questionnaire's division of stability.

To examine the relationship between pilots' subjective assessment of approach stability and the system-generated (machine-learning) assessment, presented in section 6.1, we conducted a Pearson correlation analysis. While the SADA HMI was only available live to participants in the SADA On condition, its algorithm was applied throughout all simulation exercises to assess approach stability for all scenarios, including those in the SADA Off condition. This allowed for a consistent, objective reference across conditions, enabling a comparison between human and machine evaluations regardless of real-time system availability. A positive correlation was hypothesized based on theoretical alignment between self-assessed and machine-assessed approach stability. A Pearson correlation (one-tailed) showed a small-to-moderate, **non-significant** positive association, $r(24) = 0.26$, $p = 0.098$.

To examine whether the relationship between self-assessed and machine-learning-assessed approach stability was moderated by the availability of the SADA tool, we conducted a general linear model with self-assessed stability as the dependent variable, machine-learning-assessed stability as a covariate, SADA availability (two levels: On, Off) as a factor, and their interaction as a term in the model. The overall model was **not significant**, $F(3, 22) = 0.93$, $p = 0.44$, and explained a small proportion of the variance in self-assessed stability (adjusted $R^2 = 0.00$). There was **no significant main effect** of SADA availability, $F(1, 22) = 0.01$, $p = 0.93$, **nor** of machine-assessed stability, $F(1, 22) = 0.58$, $p = 0.45$. Importantly, the interaction between SADA availability and machine-assessed stability was also **not significant**, $F(1, 22) = 0.38$, $p = 0.54$, indicating that the relationship between algorithmic and self-assessed stability did not differ as a function of whether the SADA tool was available during the approach.

Simple effects analyses showed that the relationship between machine-assessed and self-assessed stability was **not significant** in either condition (SADA On: $p = 0.46$; SADA Off: $p = 0.50$), further supporting the absence of a moderating effect.

6.3.2 Pilots' Scenario Evaluation

Figure 27 illustrates the pilots' response on how often they experience scenarios as flown in the simulation exercise in their real-world operations. The bar plot is color-coded by scenario type, summarizing scenarios one and two as standard approaches, scenarios three and four as base intercepts, and scenarios five and six as short intercepts. We observe that the short intercepts, which were designed to be the most demanding scenarios due to ATC constraints, are experienced by all pilots at least from time to time. The frequency of standard approach scenarios is not assessed uniformly by pilots, similar to the base intercepts.

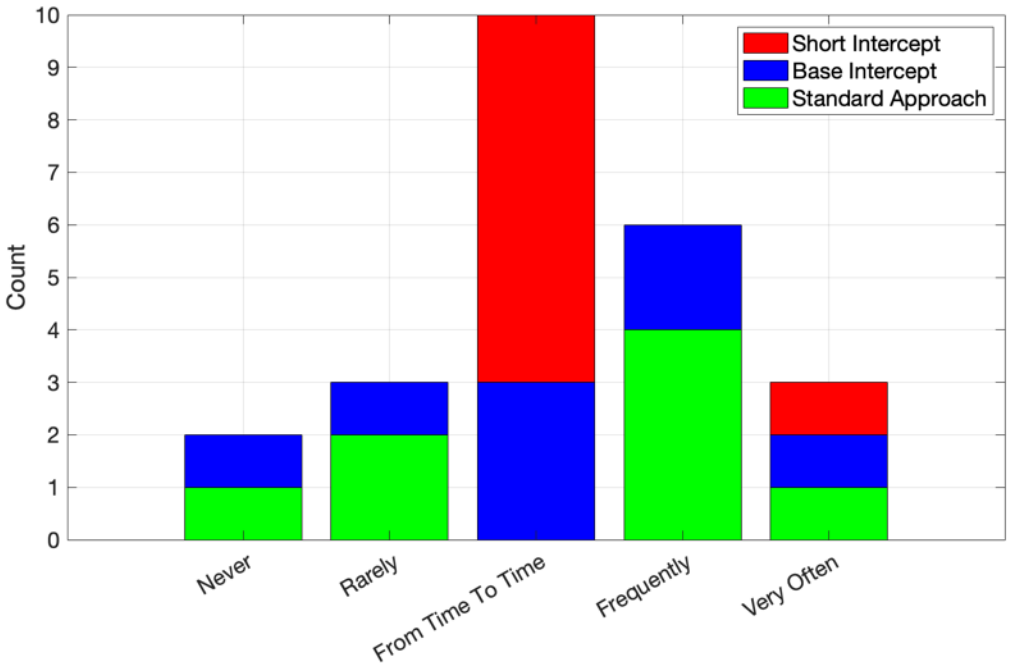


Figure 27: Pilots' Perception on the Likelihood of Similar Scenarios Occurring in Real Operation, Color Coded by Scenario Type.

A one-tailed Pearson correlation was conducted to test the hypothesis that self-assessed stability (analyzed in the previous section) decreases as perceived task difficulty increases. The result showed a small-to-moderate negative correlation that was **not statistically significant**, $r(15) = -0.24$, $p = 0.179$. While the direction of the association aligned with expectations, the result does not provide sufficient evidence to support a reliable relationship between these variables in the current limited sample.

Perceived task difficulty was rated by participants following each simulated approach and analyzed to determine whether difficulty differed across scenario types. Scenario types were grouped into three categories based on intercept geometry and operational complexity: Standard (Scenarios 1–2), Base (Scenarios 3–4), and Short (Scenarios 5–6). Given the within-subjects[4] structure of the data and small sample size, a linear mixed-effects model was used to account for repeated measures and participant-level variability.

[4] Meaning that all participants completed all scenarios in all conditions, as opposed to a 'between-subjects' design, where participants are split between conditions.

The model included **Scenario_Type** as a fixed effect and a random intercept for **Participant_ID**. Satterthwaite approximation was applied for degrees of freedom, and Wald confidence intervals were computed for fixed-effect estimates. The model was based on 18 complete observations from 4 participants, each contributing ratings across multiple scenario types.

Model fit indices suggested that a substantial portion of the variance in task difficulty was attributable to between-subject differences (conditional $R^2 = 0.487$), with the fixed effect of scenario type explaining a smaller proportion of variance (marginal $R^2 = 0.154$). The omnibus test for the fixed effect of scenario type was **not statistically significant**, $F(2, 12.3) = 2.52$, $p = 0.121$.

Estimated marginal means indicated a descriptive trend toward higher perceived difficulty in both the Base (M = 4.36) and Short (M = 4.39) scenario types compared to the Standard condition (M = 3.23), although pairwise comparisons did not reach significance (Base vs. Standard: $p = 0.068$; Short vs. Standard: $p = 0.086$). These trends are mostly consistent with scenario design expectations but should be interpreted cautiously given the small sample size and wide confidence intervals.

6.3.3 Mental Workload

As discussed in Section 5.4.3.1, the participants' subjective workload was assessed using the NASA-TLX, which includes six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Internal consistency across these items was high, with a Cronbach's alpha of $\alpha = 0.91$, supporting the use of a composite workload score. Table 13 presents the means and standard deviations for each subscale and for the overall mental workload (MWL) composite score, averaged across all conditions and scenarios. On average, participants reported higher levels of effort, mental demand, and temporal demand, with lower ratings for physical demand and frustration. The performance item was reverse-scored such that higher values indicate greater perceived performance degradation, aligning its directionality with the other NASA-TLX subscales. The composite MWL score was then calculated as the unweighted mean of all six subscales (Raw TLX). The following section presents inferential analyses examining the effects of scenario and digital assistant availability on MWL.

Table 13: Means and standard deviations for NASA-TLX subscales and composite MWL score (0–20 scale, higher = greater workload). Performance was reverse-scored to align directionality.

NASA-TLX Item	Mean (M)	Standard Deviation (SD)	Cronbach's α
Mental Demand	7.50	3.91	
Physical Demand	5.89	4.91	
Temporal Demand	6.78	2.82	
Performance (rev.)	8.06	6.08	
Effort	8.50	4.49	
Frustration	4.11	3.64	
Composite MWL	6.81	3.69	0.91

Although the Shapiro-Wilk test indicated a deviation from normality ($W = 0.89$, $p = 0.041$), visual inspection of Q-Q plots, as well as skewness and kurtosis values, suggested only mild non-normality. Given the small sample size and within-subjects structure of the data, a linear mixed-effects model was employed to account for repeated measures across participants and to allow for greater flexibility in handling distributional assumptions. This approach provides a more robust alternative to traditional repeated-measures ANOVA, particularly when assumptions of normality or sphericity may be violated.

To examine whether participants' perceived MWL differed across scenario types and SADA availability, a linear mixed-effects model was fitted using data from 4 participants, each contributing up to six trials, for a total of 24 data points. Six trials were excluded due to incomplete MWL constituent item scores (leaving a sample size of 18), resulting in a slightly unbalanced dataset. The model included **Scenario_Type** (three levels: Standard, Base, Short) and **Assistant_Availability** (two levels: On, Off) as fixed effects, with a random intercept for **Participant_ID** to account for repeated measures. Satterthwaite approximation was used for degrees of freedom, and Wald confidence intervals were computed for parameter estimates.

Model fit indices suggested a good overall fit, with a conditional R^2 of 0.855, indicating that 85.5% of the variance in MWL was accounted for when including both fixed and random effects. The marginal R^2, representing variance explained by fixed effects alone, was lower at 0.066.

The omnibus tests for the fixed effects revealed **no significant main effect** of scenario type on MWL, $F(2, 9.0) = 1.47$, $p = 0.281$, and **no significant main effect** of assistant availability, $F(1, 9.2) = 2.58$, $p = 0.142$. There was also **no significant interaction** between scenario type and assistant availability, $F(2, 9.0) = 1.11$, $p = 0.372$.

Estimated marginal means indicated a descriptive trend toward higher MWL ratings in the Short scenario type condition (M = 8.09) compared to Base (M = 6.73) and Standard (M = 6.71), although this difference was not statistically significant. Unexpectedly, MWL was numerically higher when the assistant was active (M = 7.85) compared to when it was off (M = 6.51), though again this trend did not reach statistical significance. Notably, the only condition in which the assistant slightly reduced MWL was in the Short scenario condition, where MWL dropped from 8.21 (Assistant Off) to 7.97 (Assistant On), though this difference was minimal and not reliable.

These findings suggest that, in this preliminary sample, neither scenario type nor SADA availability significantly affected perceived mental workload. However, descriptive trends may inform hypotheses for future analyses with larger samples.

6.3.4 Situation Awareness

Participants' situation awareness (SA) was measured using the SA3 questionnaire, as discussed in Section 5.4.3.1. The SA3 includes three subscales corresponding to three dimensions of SA: perception, comprehension, and projection. These three dimensions were sufficiently internally consistent, with a Cronbach's alpha of $\alpha = .80$, and thus suitable for composite representation. Table 14 presents the means and standard deviations for each constituent SA3 scale and the overall composite SA score, averaged across all conditions and

scenarios. On average, participants reported higher scores for situation comprehension than situation perception and future projection.

Table 14: Means and standard deviations for SA3 subscales and composite situation awareness (SA) score (scale: 1–11, higher = better SA). Scores are averaged across all conditions and participants.

SA3 Item	Mean (M)	Standard Deviation (SD)	Cronbach's α
Perception	8.54	1.82	
Comprehension	9.54	1.02	
Projection	8.54	2.06	
Composite SA	8.88	1.43	.80

Although all SA scores were valid, the distribution of composite SA scores showed a clear deviation from normality. The Shapiro-Wilk test was significant, W = 0.77, $p < 0.001$, and both skewness (-1.87) and kurtosis (3.59) values indicated a negatively skewed and peaked distribution, consistent with a ceiling effect in participants' responses. Visual inspection of the histogram and Q-Q plot supported this conclusion, showing clustering at the high end of the scale and systematic deviation from normality. Given the repeated-measures structure of the data and the robustness of linear mixed-effects models to violations of normality assumptions—and the exploratory nature of this early-stage dataset—we proceeded with this approach for the inferential analysis of the SA scores.

To examine whether participants' perceived SA differed across scenario types and SADA availability, a linear mixed-effects model was fitted using data from 4 participants, each of whom completed six trials (one for each combination of scenario type and assistant availability), resulting in a total of 24 observations. The model included **Scenario_Type** (three levels: Standard, Base, Short) and **Assistant_Availability** (two levels: On, Off) as fixed effects, and a random intercept for **Participant_ID** to account for repeated measures. Satterthwaite approximation was used for degrees of freedom, and Wald confidence intervals were computed for parameter estimates.

Model fit indices suggested a moderate-to-strong model fit, with a conditional R^2 of 0.422, indicating that 42.2% of the variance in SA was explained by both fixed and random effects. The marginal R^2, representing variance explained by fixed effects alone, was 0.263.

The omnibus tests for the fixed effects revealed **a significant main effect** of assistant availability, $F(1, 15.0) = 7.85$, $p = 0.013$, indicating that SA ratings were significantly higher when the assistant was available. There was **no significant main effect** of scenario type, $F(2, 15.0) = 0.82$, $p = 0.458$, and **no significant interaction** between scenario type and assistant availability, $F(2, 15.0) = 0.48$, $p = 0.626$.

Estimated marginal means showed that participants reported **higher SA** when the assistant was active (M = 9.56, SE = 0.46) compared to when it was not (M = 8.19, SE = 0.46). Descriptive differences between scenario types were minimal and non-significant: Standard (M = 8.96), Base (M = 8.46), and Short (M = 9.21). Assistant-related increases in SA were observed consistently across all scenario types, with the largest difference in the Standard condition (Assistant Off: M = 8.00; Assistant On: M = 9.92).

These preliminary results suggest that, unlike for MWL, the availability of the SADA had a statistically significant and positive effect on participants' perceived situation awareness, regardless of scenario type.

6.4 Post-Session Questionnaire

The post-session questionnaire included the System Usability Scale (SUS), described in Section 5.4.3.2, to assess overall usability of the SADA system following the simulation exercise. Responses were collected after participants had completed all six scenario runs and associated post-run questionnaires. One pilot's response was not included due to an incomplete submission.

Table 15 contains the individual SUS scores, which are 65, 77.5, and 85, yielding a mean score of 75.8. This places the system within the "good" usability range, often associated with a "B" letter grade in standardized SUS benchmarks. Figure 28 visualizes these results.

Qualitative perceptions of usability, effectiveness, and system limitations are further explored in Section 6.5, which presents a thematic analysis of post-session debrief interviews.

Table 15: System Usability Scale Results

SUS Result 1	SUS Result 2	SUS Result 3	SUS Result (Mean)
65	77.5	85	75.8

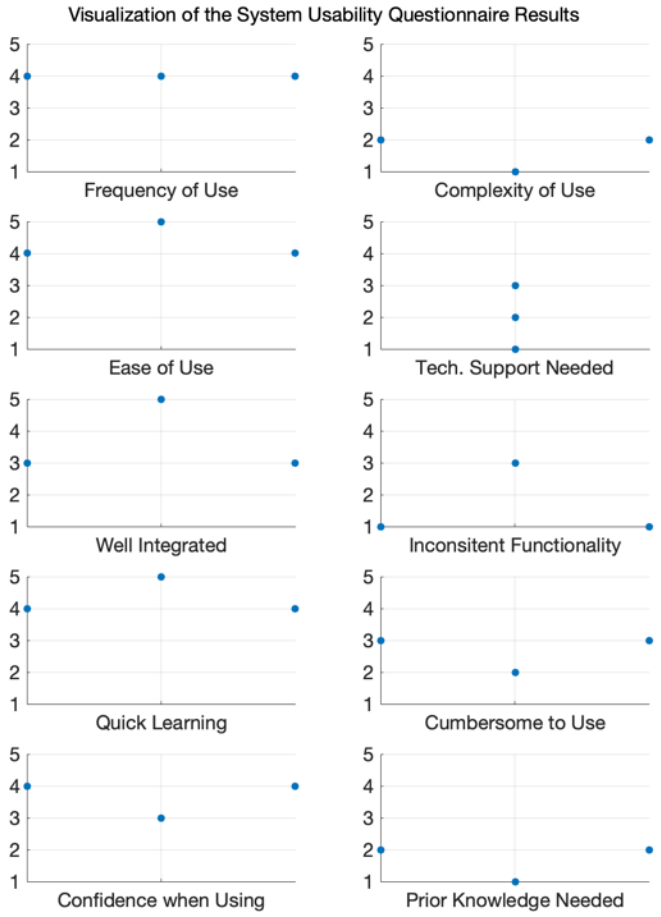


Figure 28: Visualization of the system usability scale questionnaire results as dot plots. To visualize multiple similar results, results of the same magnitude are spread horizontally.

6.5 Interview thematic findings

This section reports the qualitative findings from the interview data, analyzed as described in section 5.4.4.

Five interlocking themes capture participant pilots' expectations and experiences: **(A)** Interface & Alert Design, **(B)** Guidance Specificity & Model Transparency, **(C)** Cognitive Load & Crew Resilience, **(D)** Timing & Predictive Windows, and **(E)** Operational Context & Learning Ecosystem. Together, they illustrate how participants evaluated the SADA for its relevance, usability, and effects on cockpit operations. In the following subsections, we describe each theme in turn and conclude the section with a brief synthesis.

6.5.1 Theme A: Interface and alert design

Participants judged the Human-Machine Interface (HMI) of the Stabilized Approach Assistant in part by how clearly it caught the eye in the PFD. Comments converged on three intertwined design cues; placement, color-coding, and visual salience, all of which shaped their willingness to notice, interpret, and ultimately act on the system's messages.

- **Placement aligned with cockpit scanning habits.** Participants praised the current position "in the center of the FMA" (P1), calling it "a very prominent place in the PFD" (P1) and therefore easy to pick up in the normal instrument scan. The location was recognized as prime real estate that avoids head-down time while remaining inside the crew's focal area.
- **Color gives instant meaning; if mapped consistently.** The red/yellow coding matched long-established alert hierarchies: "It should be red for unstable and yellow for a situation where you can still be stable at the 1 000 ft gate" (P1). Pilots emphasized that sticking to these conventions prevents misinterpretation and unnecessary startle. However, pilots also valued the idea of a "traffic-light" (P1) predictor at localizer intercept—*green* (stable), *yellow* (warning; take action), *red* (alert; go around)—to increase go-around readiness.
- **Make the cue hard to miss, but not distracting.** Several suggestions targeted greater salience without adding clutter. One participant (P1) asked for the relevant data field to "blink" when it mattered, arguing that a momentary flash draws the eye faster than a static text block. Another compared the concept to Airbus' ROW/ROP runway-distance line—"a line on your runway… shows you where you can stop with maximum braking" (P2)—as an example of a succinct visual overlay that conveys urgency without words.

Taken together, these observations underline a simple principle: an advisory is only as effective as its perceptual footprint. By anchoring the alert in the crew's natural scan, coding it with universally recognized colors, and adding just-enough motion or graphical context, designers can ensure the Stabilized Approach Assistant earns pilots' attention without monopolizing it – a prerequisite for any further guidance the system may offer in later themes.

6.5.2 Theme B: Guidance specificity and model transparency

Once an alert had caught their eye, pilots judged its worth by two yardsticks: "Does it tell me what to fix?" and "Can I trust the logic behind that advice?" Their answers crystallized into two closely coupled subthemes.

Subtheme B.1 Parameter-Specific Guidance: The generic "stabilize" cue was judged a blunt instrument. Pilots asked for a pinpoint nudge toward the offending parameter so they could act without hunting for the problem: "If it says amber 'stabilize,' I would also be happy if I get a parameter which has to be stabilized … speed, glideslope, energy" (P1). Drawing a skiing analogy, P2 wanted a one-point coaching que: "If [the assistant] had said, 'concentrate on your pitch, that would have been better" '. The same participant likewise saw most benefit "if it's possible … to focus on the one variable that the user has trouble with, or the one [with the] biggest variance" (P2). Speed was singled out as the prime culprit: "I think you have your solution. It's the speed" (P2).

Participants noted that in single-pilot or high- workload situations this specificity becomes indispensable; without it, "you need three more seconds to realize which parameter is meant" (P1). Suggestions for improvement ranged from blinking the speed window or glideslope scale to a brief advisory line ("airspeed high – use drag"). Such cues, pilots argued, mirror the verbal shorthand a human PM would employ and shorten the path from detection to correction.

Subtheme B.2 Transparency and Reliability: Even pinpoint advice falls flat if its rationale is opaque or its timing unhelpful. Participants therefore probed *how* the model worked, *what data* it used, and *when* it spoke up.

- **Physics-based, energy-aware logic.** Pilots doubted a back-box classifier could outperform a forward physics calculation. P1 argued that if you used a model aware of shortcuts, wind, and weight to produce "this drag calculation, if you do this in real time, then you have a good predictive tool" (P1).
- **Data input and openness.** Participants pointed out that uplinked wind profiles are already available and could feed the prediction: "[it's possible] to have an approach wind with head and tail wind. I already have it on the AFB" (P1).
- **Right horizon, advisory tone.** A four-mile cue was too short to act upon; P2 preferred six nautical miles so that "you have 1 000 ft to change something" (P2). Given the existing 1 000-ft stabilization gate, participants stressed the alert should remain advisory, not a command: "For me, it would make more sense to have an advisory, because we already have this 1 000 ft" (P2).

Together, these insights frame a clear development target – an assistant that surfaces parameter-level cues grounded in a transparent, energy-based prediction, delivered early enough to be useful and framed as an advisory partner rather than a strict alarm.

6.5.3 Theme C: Cognitive load and crew resilience

Pilots repeatedly linked the value of the assistant to moments of high workload. During busy intercepts or hand-flown segments, even a well-placed alert can be lost if it is vague or forces additional mental search. Conversely, guidance that pinpoints *one* salient variable can free scarce cognitive bandwidth and bolster crew resilience.

- **Workload-sensitive comprehension.** When several tasks converge—"glideslope capture, flaps, overshoot, high energy as well!"—a generic "stabilize" cue costs time: "You need three more seconds to realize which parameter is meant" (P1). Those seconds matter, especially if a late "nuisance" alert turns into a scavenger hunt: "You try to check your parameters and you don't find a parameter and it might lead to exceeding another parameter" (P1).
- **Digital backup when the human PM is busy.** One pilot welcomed an assistant that "tells me whether with that energy level I'm still able to continue the approach" (P1) when the pilot-monitoring (PM) is heads-down with weather or ATC calls, stressing the need for an extra but unobtrusive set of eyes.
- **Cognitive limits and targeted coaching.** Following the skiing metaphor, you can only adjust one variable at a time: "… because you can't concentrate on 10 at the same time anyway. So, you have one variable, and that you can change. And this would be the same here" (P2). This self-awareness of recurring weak spots ("I know that I have a problem with the pitch" (P2)) underlines the appetite for concise, personalized cues rather than a scatter of marginal metrics.

A clear cross-theme link emerges here with **B.1 Parameter-Specific Guidance**: reducing the alert to *speed high* or *pitch down* not only speeds up corrective action but pre-empts cognitive overload. Vague or multiplex warnings, by contrast, risk compounding stress precisely when the pilot's working memory is taxed.

In short, resilience under pressure hinges on keeping guidance *specific* and timing it so that it *spares*, rather than *spends*, a precious crew resource – attention.

6.5.4 Theme D: Timing and predictive windows

Beyond *what* the assistant says, pilots care keenly about *when* it speaks up. They framed the tool's usefulness as a race against physics and procedures: an alert too close to touchdown is "just preparation for [a] go-around" (P1), whereas a timely cue still leaves altitude and distance to act. Two subthemes capture that temporal calculus.

Subtheme D.1 Prediction Horizon and Actionability: The current trigger—4 NM / 1 000 ft—was judged too late already to salvage energy-heavy approaches: "What the prediction is right now is more like a preparation for go-around. If you're not stabilized at 4 miles, it's, yeah" (P1). P2 echoed this, arguing for a 6-NM window: "Six miles, I would say. So, you have 1 000 ft to change something" (P2). Participants saw value in matching Airbus ROW/ROPs logic; a hard-stop warning at 500 ft when elevator authority and ground-speed protections limit further corrections: "So if you are not stable in 500 ft, they say, 'okay, this is the last line,' because they don't think that you can make this situation much better after 500 ft" (P2).
Taken together, pilots proposed a tiered timeline: an early advisory at ~6 NM to prompt drag or configuration changes, followed (if needed) by a non-negotiable go-around trigger near 1 000 ft.

Subtheme D.2 Nuisance Alerts and Trust Erosion: When the cue arrives late, or triggers despite operational constraints, it not only fails to help – it can erode confidence.

- **Nuisance effect under load.** P1 described scanning for a fault after an amber "stabilize" cue and finding none: "If you are in a situation where you have a lot of things going on at the same time and you get a nuisance stabilize indication, that would be probably the worst case. Then you try to check your parameters, and you don't find a parameter and it might lead to exceeding another parameter" (P1).
- **False hope and risk-taking.** A late stabilize call can "trigger you like 'I can still make it'" (P1), encouraging futile tinkering instead of an earlier go-around decision.
- **Physical limits.** Below 500 ft the flight-control system gradually restricts elevator authority, making large pitch-or-speed corrections impossible: "you only have a short time" (P2).
- **Procedural limits.** Controllers often clear jets to hold 170 kt until 4 NM, an ATC practice airline accept in their manuals; any alert must recognize that "later stabilization in airspeed may only be acceptable for ATC procedures or instructions" (P2).

When prediction timing and operational reality diverge, crews start to ignore the system—or worse, fight it. Conversely, calibrating the horizon to give *real maneuvering* margin (Subtheme D.1) directly reduces these nuisance triggers, reinforcing the trust loop highlighted in the second theme (section 6.5.2). In sum, time is the currency of action: give pilots enough of it, and the assistant becomes a valued teammate; squander it, and the message turns into noise.

6.5.5 Theme E: Operational context and learning ecosystem

This theme shifts the lens from the interface itself to the broader operational and learning environment in which the Stabilized Approach Assistant would exist. Participants stressed that a prediction tool can only succeed if it respects the real-world constraints that generate unstable approaches and (equally importantly) if its outputs are recycled into personal and organizational learning loops. Two subthemes emerged:

Subtheme E.1 External Operational Pressures: Pilots identified a recurring set of factors outside the cockpit that routinely load the aircraft with excess energy before the 1000 ft gate:

- Eco-flying incentives. In the name of fuel burn and noise, crews may delay flap or gear: "In terms of sustainability, you try to set as less flaps as possible with the glideslope … which might lead up in a high energy state of 1 000 ft" (P1).
- Controller-imposed speeds. Pilots described occasions when ATC asked them to "keep 170 kt until four miles," leaving too little altitude to decelerate before the 1000-ft stabilization gate. As P2 noted when recalling a live operational event: "You can, but of course we have to be stabilized at 1 000 ft. So [the co-pilot] said to me, 'we can't do it'" (P2). He later pointed to the airline's Operations Manual, observing that "a later stabilization in airspeed may only be acceptable [if mandated by] ATC procedures or instructions" (P2), illustrating how crews weigh external demands against their own safety SOPs. P1 echoed the practical effects of such clearances: "From [an] operational perspective, often you get shortcuts with less track miles and that's why you're intercepting your localizer with high energy, for example" (P1).

These excerpts reinforce the idea that externally imposed speed or track-mile constraints are a routine, systemic source of "extra energy" on final, and therefore a prime target for any predictive guidance or post-flight learning the Stabilized Approach Assistant aims to deliver.

Subtheme E.2 Post-flight Analytics and Evidence-Based Training: Beyond the heat of the approach itself, pilots framed the Stabilized Approach Assistant as a *post-flight tutor;* a source of replay data that can transform one unstable episode into actionable practice goals for the next. Pilots saw their greatest long-term value in replaying the Stabilized Approach Assistant's data after a flight to isolate the specific "bars" they should practice before the next "performance." This targeted feedback turns a single unstable approach into evidence-based training material. P2 compared the desired feedback to an instrumental-music coach who intervenes well before the final performance: "It's like you have … a concert … on the 30th of September… if [a] teacher [listens] … at the end of August … and [says], *'This is the part you have to practice… only concentrate on those three.'* … Those three pieces are my pitch and my speed, maybe. So, for me it's more a training issue" (P2). The analogy underlines the value of targeted, time-shifted coaching: rather than a generic "you were unstable," pilots want granular, parameter-specific insights that can be reviewed shortly after the flight and turned into concrete practice goals for the next: "I do this five times, that's evidence-based training. You look at me and you say, 'Hey, [P2], your pitch was not so good.' This is evidence-based training" (P2).

6.5.6 Integrative discussion

Taken as a whole, the five themes tell a consistent story: pilots welcome a digital "crewmember" that spots developing instability early, pinpoints the culprit parameter(s), and blends seamlessly with existing cockpit systems and cues—but only if the system earns trust by matching crew workflow and real-world constraints. **Utility** hinges on *timeliness* (Theme D): a tiered horizon—advisory at ~6 NM, hard stop near 1000 to 500 ft—gives crews genuine maneuvering margin and curbs nuisance alerts that erode confidence (Themes B and D). **Usability** begins with perceptual design (Theme A): a traffic-light color code, prime PDF placement, and minimal but attention-grabbing motion to ensure that the message is noticed without adding clutter. Yet attention is only half the battle; cognitive economy

(Theme C) demands minimum- or single-parameter guidance, so pilots are spared a scavenger hunt at peak workload.

Equally important is **explainability**. Participants linked trust to a transparent, physics-based prediction model and to clear data inputs (wind, weight, drag) rather than opaque probability scores (Theme B). When the assistant's logic mirrors the way pilots' reason themselves—solving the "energy differential"—they are more inclined to accept its advice (i.e., there is an alignment in mental models), whether as an in-the-moment prompt or as post-flight evidence for targeted practice (Theme E). Re-playing specific pitch- or speed-related "bars" in training settings after a flight turns isolated alerts into concrete learning goals, completing a virtuous loop from real-time aid to evidence-based training.

Based on these limited and preliminary qualitative results, the Stabilized Approach Assistant is judged **promising but conditional**. Its value materializes when four design commitments are met simultaneously:

1. perceptually salient yet unobtrusive interface,
2. parameter-specific, workload-sensitive guidance,
3. prediction windows aligned with physical and procedural limits, and
4. transparent logic that feeds a broader learning ecosystem.

Fulfilled together, these facets position the assistant as a trusted, pro-active teammate rather than another cockpit warning system.

7 Conclusion and Outlook

This section interprets the results presented in Section 6 with respect to the research questions posed in Section 4. Therefore, the following subsection discusses each research question, based on the relevant data, as defined in Table 3. Based on the conclusions drawn, the final subsection covers the outlook by discussing the necessary modifications to the concept and potential next steps in the development of the SADA system.

7.1 Addressing the Research Questions

This section addresses, answers, and discusses each research question (RQ1—5) in turn.

7.1.1 RQ1: What is the relationship between pilots' perception of approach stability and the SADA's perception of approach stability?

Although the quantitative data did not reveal a statistically significant correlation between pilots' self-assessed stability ratings and the SADA's algorithmic assessments, a small positive trend was observed (see Section 6.3.1). This tentative alignment suggests a degree of shared judgment between pilots and the SADA (particularly in clear-cut cases) but also highlights divergence in more ambiguous scenarios.

Interview findings reinforce this picture. Pilots' trust in the SADA's stability judgments hinged on how clearly its cues matched their own mental models of approach dynamics. Specifically, participants stressed the need for parameter-specific feedback (Theme B.1 Parameter-Specific Guidance), such as alerts tied directly to excess speed or pitch deviations, to understand and validate the system's outputs. Without this specificity, even a technically accurate prediction risked being perceived as vague or unhelpful.

This perception gap may stem in part from how the machine learning model was trained. As noted, the SADA's classification logic reflects formal instability definitions derived from flight data monitoring (FDM) thresholds. While such labeling supports reliable detection of pronounced instability, it may limit the model's sensitivity to borderline or recoverable cases, i.e., precisely the situations where pilots reported most valuing timely guidance. In essence, the system's precision-focused training may skew its outputs toward "obvious" instabilities, which pilots already recognize unaided.

To better support human decision making, future model development should prioritize the identification of preventable unstable approaches; those that could be stabilized with pilot action following an early cue. Achieving this would likely require augmenting the training dataset with dynamic performance metrics (e.g., energy state, drag potential, or aircraft-specific constraints) and revisiting the labeling logic to reflect not just outcomes but also the recoverability horizon at the time of prediction.

While the current system shows early signs of perceptual alignment with pilot assessments, deeper integration of pilot reasoning and cockpit context (both in interface design and model development) will be key to bridging the gap between algorithmic judgment and operational trust.

7.1.2 RQ2: How does the SADA affect pilots' situational awareness during final approach?

The preliminary quantitative findings indicate a statistically significant and positive effect of SADA availability on pilots' self-reported situation awareness (SA), as measured by the SA3 scale (Section 6.3.4). This effect was consistent across all scenario types, suggesting a general benefit of the assistant in supporting pilots' awareness of approach dynamics during final approach. This consistency is notable, given that pilots rated the Base and Short intercept scenarios as more difficult than Standard scenarios (see Section 6.3.2). These difficulty ratings aligned with the intended scenario design and underscore that the observed SA benefits were present even under higher operational complexity. While sample size limitations and ceiling effects in the data prevent strong generalization, these results tentatively support the assistant's design objective of enhancing SA in the critical final approach window.

Importantly, the version of the SADA used in this study provided only binary, non-specific guidance by flagging an approach as stable or unstable without identifying which parameters contributed to that status. Despite this limitation, participants still reported increased SA when the system was available, suggesting that even coarse-grained feedback may help prime pilots' attention toward potential deviations or reinforce their own assessments.

However, interview data clearly point to the potential for greater SA support through more specific, transparent guidance. As captured in Theme B (Guidance Specificity and Model Transparency), participants consistently expressed a desire for cues that identify which parameter (e.g., speed, pitch, vertical path) was trending toward instability. Pilots noted that such specificity would allow them to move more quickly from detection to correction, especially in high-workload situations; an observation echoed in Theme C (Cognitive Load and Crew Resilience). This kind of targeted support could enhance not only awareness of the current state (perception and comprehension) but also forward-looking judgment (projection), by clarifying the likely trajectory of the approach.

While the current SADA implementation appears to positively influence SA, its impact may be limited by the generic nature of its cues. The qualitative findings suggest that future iterations offering parameter-specific, explainable alerts could further enhance pilots' situational understanding (particularly under time pressure) thereby advancing the assistant's core aim of supporting timely, informed decision-making in complex operational contexts.

7.1.3 RQ3: How does the SADA affect pilots' mental workload during final approach?

The quantitative results did not show a statistically significant effect of either SADA availability or scenario type on self-reported mental workload (MWL), as measured by the NASA-TLX (Section 6.3.3). While there was a descriptive trend toward higher MWL ratings when the assistant was available, this pattern was not reliable and varied across scenario types. In the "Short" scenario (a condition with inherently elevated time pressure), the assistant slightly reduced MWL on average, though this reduction was modest and

statistically insignificant. Importantly, independent ratings of scenario difficulty confirmed that pilots perceived the Short and Base scenarios as more demanding than Standard approaches (Section 6.3.2). This supports the scenario design assumptions and suggests that the non-significant MWL variations across conditions reflect real differences in task complexity. Overall, these preliminary results suggest that the presence of the SADA did not introduce a measurable increase or decrease in perceived workload.

From a design perspective, the absence of a significant MWL increase is nonetheless encouraging. A key goal of the assistant was to support decision making during final approach without adding to the crew's cognitive burden. These early findings suggest that the current implementation of the SADA, while basic in its guidance, does not appear to overload the pilot during the high-stakes final approach phase.

However, the qualitative findings reveal a more nuanced picture. Theme C (Cognitive Load and Crew Resilience) from the interview data captures pilots' reflections on how SADA could influence workload (both positively and negatively) depending on its design. In particular, participants indicated that vague or non-specific alerts (such as the current binary "stabilize" cue) can be counterproductive under high workload. When pilots are already managing multiple concurrent tasks (e.g., flaps, glideslope, ATC instructions), a generic alert can trigger a "scavenger hunt" for the root cause, momentarily increasing mental effort. As one participant noted, "you need three more seconds to realize which parameter is meant," and time may not be available during a fast-evolving approach segment.

Conversely, pilots described how parameter-specific guidance would offload cognitive work by pointing directly to the most relevant issue (e.g., excess speed), allowing for faster and more confident correction. This form of "targeted coaching" was considered especially valuable in scenarios involving high tempo or limited monitoring capacity, for example, when the pilot flying is heads-down or the pilot monitoring is engaged with external tasks.

These insights suggest that while the current implementation of SADA does not appear to elevate workload, its potential to *actively reduce* MWL may be limited by its lack of specificity. Future refinements aimed at minimizing cognitive overhead, particularly under time-constrained or high-load conditions, could include single-parameter alerts, more intuitive visual cues, or integration with task-relevant monitoring logic.

The assistant's current design seems workload-neutral, meeting its design aim of avoiding undue burden, but greater gains in workload relief may be achievable through more precise and context-sensitive alerting strategies.

7.1.4 RQ4: What is the effect of the SADA on go-around compliance?

Due to the limited number of unstable approaches and an even lower incidence of actual go-arounds, a quantitative analysis of go-around compliance was not feasible within the scope of this simulation study. Only one recorded go-around occurred during an approach that was flagged as unstable by the assistant. In this case, the pilot noted in the post-scenario questionnaire that the decision to go around was already forming but that the SADA "helped to [ensure] and execute the go-around" (P1). This comment suggests that the assistant may

act as a reinforcing cue in marginal or high-uncertainty cases, helping pilots align their actions with standard operating procedures. Notably, pilots rated the more demanding Short and Base intercept scenarios as representative of situations they encounter in real operations (Section 6.3.2), adding ecological validity to both the observed go-around and the broader set of qualitative insights gathered during these higher-complexity runs.

Qualitative interview data reinforces this interpretation. Participants consistently indicated that a tool like SADA could improve go-around compliance by legitimizing the decision to discontinue an unstable approach. As discussed in Theme D (Timing and Predictive Windows), pilots stressed that late or vague cues are less useful and may even encourage "hopeful" continuation. However, when alerts are delivered with sufficient time and clarity, they can shift the pilot's mindset from salvage to decision execution, which is particularly important in high workload or ambiguous situations.

Further, in Theme B.2 (Transparency and Reliability), participants discussed that go-around decisions are easier to justify when the system's logic is understandable and mirrors pilot reasoning. A model that surfaces energy-related instability early and does so with credible cues (e.g., based on wind, drag, or configuration state) can help confirm a pilot's intuition and reduce hesitation. In this sense, the assistant is not just a detector but becomes a cognitive ally that supports confident, timely compliance with safety protocols.

Pilots also viewed the assistant's value through the lens of post-flight learning. As highlighted in Theme E (Operational Context and Learning Ecosystem), the SADA's potential to generate reviewable data could reinforce go-around criteria through evidence-based training. Even in cases where a go-around is not executed, reviewing why an alert occurred could strengthen decision-making strategies in future flights.

While the current dataset does not allow firm conclusions about the effect of the SADA on go-around rates, early qualitative evidence suggests that the assistant can positively influence go-around compliance, not by forcing decisions, but by reinforcing pilots' situational judgment in real-time and after the fact. Future studies with larger samples and more go-around scenarios will be needed to quantify this effect more rigorously.

7.1.5 RQ5: What is the perceived usability of the SADA system?

Post-session questionnaire responses indicate that the SADA system was generally perceived as usable, receiving scores in the "good" range on the SUS scale (Section 6.4). While limited by sample size, this early evidence suggests that the assistant's basic design concept and implementation align well with pilots' expectations and cockpit conventions. This positive assessment of usability was given after pilots had completed a range of scenarios (including Short and Base intercepts) that were rated as more demanding yet operationally realistic (Section 6.3.2). This cautiously suggests that the assistant's usability held up even in contexts that imposed greater task complexity, enhancing the credibility of the overall rating.

Qualitative feedback from the post-session interviews provides richer insight into the perceived usability of the system. Pilots appreciated the assistant's integration into familiar cockpit workflows, especially the use of established alert hierarchies, intuitive color coding, and central placement on the PFD, as captured in Theme A (Interface and Alert Design).

These design features were discussed as contributing to perceptual clarity and ease of use without adding unnecessary distraction.

However, usability was also shaped by how effectively the system communicated its reasoning. Several pilots requested that the SADA provide more parameter-specific guidance, noting that a generic "stabilize" cue could prompt unnecessary cognitive effort, particularly under time pressure (see Theme B.1 and Theme C). Some suggested that indicating the problematic parameter (e.g., speed or vertical path) could help confirm the validity of an alert or dismiss it more quickly in the case of false positives. Notably, one pilot who experienced a likely nuisance alert stressed this point, arguing that specific feedback would have enabled faster diagnosis and resolution.

That said, opinions varied: others expressed concern that increased detail might clutter the display or distract from primary flight tasks. This tension highlights a broader usability trade-off between simplicity and explainability. A potential solution, raised both in interviews and prior design discussions, is to offer pilot-selectable detail, allowing additional information to be toggled on demand. While promising, such an approach would need to be carefully evaluated to avoid undermining the interface's current strengths.

A second recurring usability concern involved the timing of the system's prediction. Pilots frequently noted that the 4 NM cue came too late to allow meaningful corrective action during high-speed approaches, a point captured in Theme D.1 (Prediction Horizon and Actionability). In contrast to the user research data collected for Deliverable D3.2 (SafeTEAM) which informed the specific implementation of the SADA, current participants proposed moving the advisory point to ~6 NM or linking it to localizer intercept or height above field elevation. These suggestions reflect a desire for cues that align better with actual maneuvering margins, thus highlighting a key intersection between usability, trust, and operational relevance.

Finally, these findings echo conclusions drawn under RQ1, where model training was identified as a limiting factor in the assistant's current utility. If the SADA is to better support preventable instability, both the machine-learning model and its prediction trigger logic may need to be reoriented toward earlier, actionable thresholds, even if this comes at the cost of lower precision. Ensuring that such changes do not degrade usability will require ongoing balancing between information richness, timing accuracy, and interface simplicity.

The SADA's early usability is promising, grounded in clear and familiar design principles. However, as pilots' feedback makes clear, continued refinement in timing, parameter feedback, and customizability will be important to ensure that usability is preserved even as functional complexity grows.

7.2 Outlook

The study presented in this deliverable report marked a significant step in the maturation of the Stabilized Approach Digital Assistant (SADA). Building on the Unstable Approach Prediction Model developed under SafeClouds.eu, Task 4.2 of the SafeTEAM project successfully integrated and evaluated the concept in a high-fidelity research simulator,

demonstrating technical feasibility in a relevant environment. As a result, the system's Technology Readiness Level (TRL) advanced from 4 to 6.

Beyond this milestone, the study provided encouraging evidence that the SADA can support key safety and decision-making goals. While quantitative data remained limited, the assistant was associated with improved pilot SA across diverse scenario types and was perceived as usable and well-integrated with cockpit workflows. In the one observed go-around, the system was credited with reinforcing the pilot's decision, and interview feedback suggests that the SADA could meaningfully support go-around compliance by legitimizing action in high-uncertainty cases.

Importantly, scenario difficulty ratings confirmed that the Short and Base intercepts used in the study were perceived as realistic and demanding, lending ecological validity to both the pilot feedback and system performance under test conditions. That usability and perceived SA benefits held up under these conditions reinforce the assistant's practical potential.
At the same time, the study identified clear areas for improvement. While the current system issues its advisory at 4 NM, this was often perceived as too late for effective intervention, especially in high-speed approaches. Pilots recommended an earlier cue: at 6 NM, at localizer intercept, or based on height above field elevation. Supporting such changes will likely require updates to the Unstable Approach Prediction Model, particularly its training and labeling strategy. Incorporating physically informed labels that distinguish between preventable and unpreventable instability could not only improve predictive utility but also align the system more closely with pilot reasoning, thereby improving trust.

The study also highlighted a potential enhancement to the HMI. Although the interface was generally well-received for its familiarity and salience, one observed nuisance alert underscored the need for optional parameter-specific feedback. Making the system's underlying reasoning visible, such as highlighting speed or glidepath as the cause for an alert, could reduce cognitive load and improve diagnostic efficiency. While earlier discussions had rejected this feature, the current results suggest that a pilot-selectable reasoning display might offer a practical compromise between simplicity and explainability.

Participants identified long-term value in the assistant's role as a post-flight learning tool. By enabling evidence-based reflection on approach dynamics and stability cues, the SADA could contribute not only to in-the-moment support but also to a broader learning ecosystem to reinforce safety margins over time. While we feel that additional tests in the simulator to gather more pilot feedback are needed, and indeed *must* be done, there is sufficient evidence to support the findings in this report. Conservatively, the TRL of assisted landings has been taken from TRL4 to TRL6. An airplane manufacturer or supply chain provider would be an ideal recipient of the technology developed in this publicly funded project.

In summary, these findings point to a clear path of development. With refinements to prediction timing, training methodology, and interface flexibility, the SADA can evolve into a proactive, pilot-aligned assistant that not only detects instability but also helps prevent it.

8 References

Annex 6: Operation of Aircraft, Part I - International Commercial Air Transport - Aeroplanes: to the Convention on International Civil Aviation (Twelfth Ausg., Bd. Annex 6). (2022). Montreal, Quebec: International Civil Aviation Organization.

Audacity. (kein Datum). Von https://www.audacityteam.org/ abgerufen

Braarud, P. Ø. (2021). Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work. *International Journal of Industrial Ergonomics, 86*, S. 103233.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), S. 77–101.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), S. 328–352.

Brooke, J. (kein Datum). SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry* (S. 207–212). CRC Press.

Data4Safety. (2022). *Guidance for identifying unstable approaches with flight data*. (Data4Safety, Produzent) Abgerufen am 19. 3 2025 von https://www.easa.europa.eu/sites/default/files/dfu/d4s_-_guidance_for_identifying_unstable_approach_with_flight_data_v3.pdf

DG Digit. (kein Datum). *EUSurvey*. Abgerufen am 26. 6 2025 von Create better online surveys and forms: https://ec.europa.eu/eusurvey/home/welcome

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), S. 32–64.

European Comission. (2012-10-05). Regulation (EU) No 965/2012.

Flight Safety Foundation. (2000). *Approach and Landing Accident Reduction Tool Kit*. Abgerufen am 19.03.2025 von FSF ALAR Briefing Note 7.1 - Stabilized Approach: https://flightsafety.org/wp-content/uploads/2016/09/alar_bn7-1stablizedappr.pdf

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), S. 904–908.

Hart, S. G., & Staveland, L. E. (kein Datum). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research.

International Air Transport Association. (2014). *Runway Excursion Statistics*. (International Air Transport Association, Produzent) Abgerufen am 19. 03 2025 von https://www.icao.int/MID/Documents/2014/MID-RRSS-2/S3%20P1%20Runway%20Excursion%20Statistics.pdf

International Air Transport Association. (2017). *Unstable Approaches*. (International Air Transport Association, Produzent) Abgerufen am 19. 3 2025 von Risk Mitigation Policies, Procedures and Best Practices: https://www.iata.org/contentassets/7a5cd514de9c4c63ba0a7ac21547477a/iata-guidance-unstable-approaches.pdf

Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction, 34*(7), S. 577–590.

Martinez, D., Fernández, A., Hernández, P., Cristóbal, S., Schwaiger, F., Nunez, J., & Ruiz, J. (2019). Forecasting Unstable Approaches with Boosting Frameworks and LSTM Networks.

Metcalfe, R. (kein Datum). *OpenVINO™ AI Plugins for Audacity.* Abgerufen am 23. 06 2025 von https://github.com/intel/openvino-plugins-ai-audacity/blob/main/doc/feature_doc/whisper_transcription/README.md

Paoli, S. (2024). Further Explorations on the Use of Large Language Models for Thematic Analysis. Open-Ended Prompts, Better Terminologies and Thematic Maps. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 25*.

SafeClouds Consortium. (kein Datum). *SafeClouds.eu*. Von https://cordis.europa.eu/project/id/724100/de abgerufen

SafeTEAM. (kein Datum). Human Factors Design Principles for a Stabilised Approach Digital Assistant. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5fdfcee93&appId=PPGMS.

Sauro, J., & Dumas, J. S. (kein Datum). Comparison of three one-question, post-task usability questionnaires.

Uzun, M. (2024). *Model Based Testing of an Unstable Approach Prediction Algorithm.*

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2025). Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans, 4*, S. 100144.

Appendix A Data Collection Material

A.1 Post Scenario Questionnaire

Post-Scenario Questionnaire

* **Please select the Scenario**

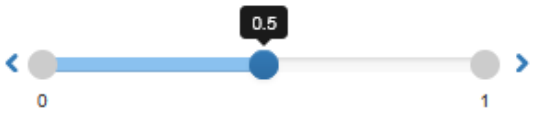
Was the Digital Assistant Activated?

- ○ Yes
- ○ No

Stability Self-Assessment

How would you rate your approach stability in this scenario?
Consider values from 0 to 0.4 as unstable, 0.6 to 1 as stable, and 0.4 to 0.6 as a grey zone.
Move the slider or accept the initial position.

Completely Unstable — Completely Stable



If self-assessed below 0.6 (not stable); which of the following factors contributed most to instability?

- ☐ Speed
- ☐ Glidesplope Deviation
- ☐ Sinkrate
- ☐ Flap Configuration
- ☐ Gear Configuration
- ☐ Localizer Deviation
- ☐ Other (please specify in text field below)

Other factors contributing to instability:

Please briefly explain the primary reason(s) for your rating above.

Please briefly explain the primary reason(s) for your rating above.

Scenario Difficulty

Overall, this task was?

Move the slider or accept the initial position.

Very easy Very difficult

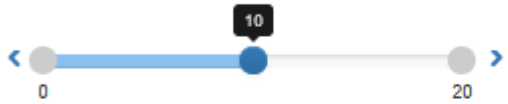
* **How often does something comparable happen to you during work?**

- ◯ Never
- ◯ Rarely
- ◯ From time to time
- ◯ Frequently
- ◯ Very often

Workload

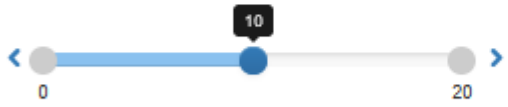
How mentally demanding was the task?
Move the slider or accept the initial position.

Very Low Very High



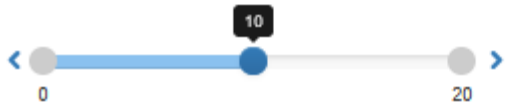
How physically demanding was the task?
Move the slider or accept the initial position.

Very Low Very High



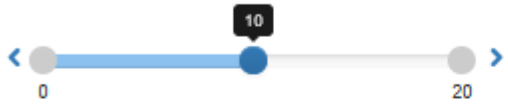
How hurried or rushed was the pace of the task?
Move the slider or accept the initial position.

Very Low Very High



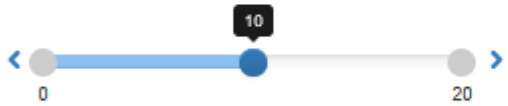
How successful were you in accomplishing what you were asked to do?
Move the slider or accept the initial position.

Perfect Failure



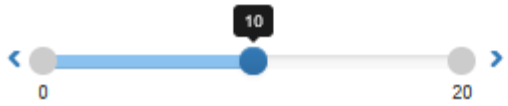
How hard did you have to work to accomplish your level of performance?
Move the slider or accept the initial position.

Very Low Very High



How insecure, discouraged, irritated, stressed, and annoyed were you?
Move the slider or accept the initial position.

Very Low Very High



A.2 Post Session Questionnaire

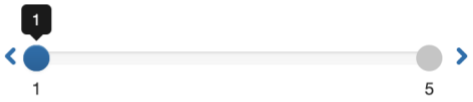
SafeTEAM Post Session Questionnaire

Usability

I think that I would like to use this system frequently.

Move the slider or accept the initial position.

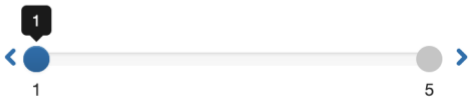
Strongly disagree — Strongly agree



I found the system unnecessarily complex.

Move the slider or accept the initial position.

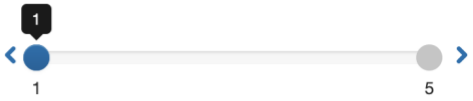
Strongly disagree — Strongly agree



I thought the system was easy to use.

Move the slider or accept the initial position.

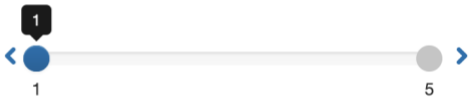
Strongly disagree — Strongly agree



I think that I would need the support of a technical person to be able to use this system.

Move the slider or accept the initial position.

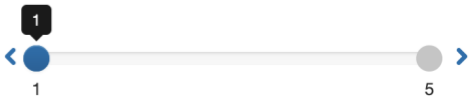
Strongly disagree — Strongly agree



I found the various functions in this system were well integrated.

Move the slider or accept the initial position.

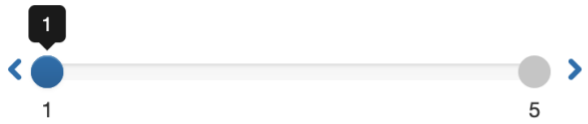
Strongly disagree — Strongly agree



I thought there was too much inconsistency in this system.

Move the slider or accept the initial position.

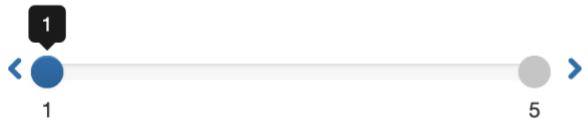
Strongly disagree — Strongly agree



I would imagine that most people would learn to use this system very quickly.

Move the slider or accept the initial position.

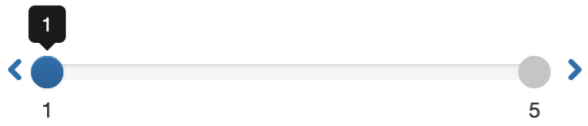
Strongly disagree — Strongly agree



I found the system very cumbersome to use.

Move the slider or accept the initial position.

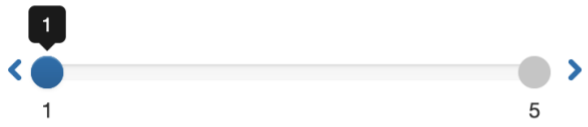
Strongly disagree — Strongly agree



I felt very confident using the system.

Move the slider or accept the initial position.

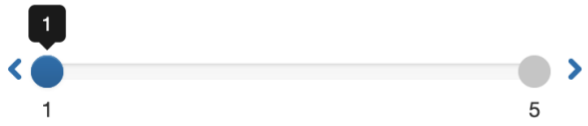
Strongly disagree — Strongly agree



I needed to learn a lot of things before I could get going with this system.

Move the slider or accept the initial position.

Strongly disagree — Strongly agree



Submit

A.3 Post-Session Interview Guide

A.3.1 General Recommendations

- Use Prompts and Follow-Up Questions: If participants give short or vague answers, encourage them to elaborate by asking questions like:
  "Could you tell me more about that?" or "What made you feel that way?"
- Ask Open-Ended Questions: Encourage richer, more detailed responses by avoiding yes/no questions.
- Prioritize Key Topics: If time is limited, focus on your primary research questions, e.g., Workload, Communication, Decision-Making, and Trust. The rest topics are optional if time allows.
- Encourage Specific Examples: Ask participants to describe concrete situations to anchor responses in concrete experiences. For example:
  "Could you walk me through what happened in that scenario?"

A.3.2 Questions

A list of questions to guide the semi-structured interview, organized by topics / dependent variables.

Communication

- Q1: "After the DA message appeared, did you notice any changes in how or when you communicated — either with your co-pilot or with ATC?"
- Q2: "Do you think the DA's alert or indication prompted you to make certain callouts sooner, later, or in a different way than usual?"
- Q3: "Can you recall a specific instance where the DA notification influenced the conversation or coordination between you and your co-pilot?"
- Follow-up: "Did you find that helpful or distracting in any way?"

Teamwork and Coordination

- Q1: "How, if at all, did the DA alter the way you and your co-pilot divided tasks or responsibilities?"
- Q2: "Did the DA ever cause confusion about who should do what, or did it help clarify your next steps?"
- Q3: "If you imagine this DA in long-term use, how do you think it might influence (strengthen or weaken) teamwork between pilots?"
- Possible prompt: "Could it inhibit your usual crew resource management practices, or might it help reinforce them?"

Workload

- Q1: "Overall, would you say the DA increased your mental workload, reduced it, or didn't really affect it? Can you share a specific example?"
- Q2: "Were there moments when you felt the DA relieved you from certain tasks, or did it add more tasks (e.g., verifying or double-checking its alert)?"
- Q3: "When you rated your workload after each scenario, what main factors came to mind? (Probe for perceived complexity, time pressure, etc.)"

Situational Awareness

- Lower priority, but may yield interesting anecdotes
- Q1: "Did the DA message affect, in any way, how you understood what was going on during the approach?" (If yes: "Could you describe how?")
- Q2: "Did the DA help you predict future conditions (e.g., a potential unstable approach), or did you still mostly rely on your own sense of the aircraft's status?"
- Q3: "Can you recall a time the DA highlighted something you hadn't already noticed? How did that shape your awareness or next steps?"

Problem-Solving & Decision-Making

- Q1: "Did having the DA change how you approached certain decisions — for example, deciding whether to continue or go around?"
- Q2: "Did you find yourself using other cockpit instruments or resources differently when the DA was present?"
- Q3: "Thinking long-term — do you see any risk that relying on such a tool like this could lead to pilots gradually losing some skills or expertise over time? Why or why not?"
- Q4 (if time allows): "Would you say the tool strengthened or reduced your sense of ownership over problem-solving?"

Usability

- Q1: "How would you describe the overall usability of the DA in its current form (e.g., clarity, placement, timing of the message)?"
- Q2: "Were the visual cues (color, font, label) easy to notice in the Primary Flight Display, or did anything feel hart to spot or distracting?"
- Q3: "How did you feel about the timing of the DA alert? Did it usually feel to early, too late, or about right for your decision-making during approach?"
- Q4: "If you could change one thing about how the DA's message or interface is displayed, what would you change?"

Human-Autonomy Teaming

- Q1: "Before using the DA, what level of 'intelligence' or interactivity did you expect from it? Did your experience match that expectation, or was it different?"
- Q2: "In an ideal future, would you prefer a future version of the DA to function more like a 'teammate' (providing suggestions, reasoning, etc.) or would you rather it stay more like a simple alert? Why?"
- Q3: "Are there other ways the DA could get your attention — like modalities (audio signals, haptic feedback) or forms of feedback? Would those feel helpful, or risk being distracting in the cockpit?"

Trust (Reliability & Agreement)

- Q1: "Which factors contributed most to how much you relied on or trusted the DA? (E.g., perceived accuracy, your own experience, scenario difficulty)"
- Q2: "Did the DA seem consistent across different scenarios? Any scenario where you found it less trustworthy?"
- Q3: "When you disagreed with the DA's assessment, how did you reconcile that discrepancy? Did you completely ignore it, or double-check your own metrics?"
- Q4 (if relevant): "If you had more time or more scenarios with the DA, do you think your trust or reliance on it would change?"

Appendix B Pilot Information Leaflet / Consent Form

B.1 Overview

SafeTeam, "**Safe** Human-digital assistant **Team**ing in the advent of higher levels of automation in aviation", is an Innovation Action (IA) funded by the European Commission under HORIZON-CL5-2021-D6-01-13. The action is coordinated by Innaxis; the Consortium further includes AESA (Spanish Aviation Safety Agency), Technical University of Munich, DataBeacon, ONERA, RISE (Research Institutes of Sweden), Pegasus Airlines, and UK CAA International as associated party.

The project was launched in June 2022 and is now entering its final phase, the simulator evaluations. In the preceding work, we designed and implemented a machine learning-based, unstable approach prediction system into the Institute of Flight System Dynamics' research simulator. In the planned simulator exercises for the Unstable Approach Case Study, we aim to evaluate this system based on pilots' feedback.
Therefore, we plan to:

- Perform simulator exercises with airline pilots, flying several approach scenarios with and without the developed assistant system
- Evaluate the approaches based on pilot feedback and the machine-learning algorithm
- Compare the pilots' perspectives with the machine-learning results to understand if there is a common situational assessment of the approaches
- Collect pilot feedback on their experience in the simulator and how the assistant is designed

In the following, we provide a more detailed explanation of the simulation exercise and the simulator itself.
The goal of the simulation exercise is to evaluate the machine learning-based, Stabilized Approach Assistant. Therefore, we will record data before, during, and after the simulator exercises. You will find a detailed explanation of which data we will record, as well as a Consent Form, which participants need to sign if they want to participate in the simulator exercise.
In addition to the data we need to record, we would like to video record the simulation exercise as redundancy in case we miss something in the protocol/minutes. We are aware that some people are not comfortable with video recordings. Therefore, we will only video record the simulation exercise if you consent explicitly / opt-in. The details are explained in the section: Data we will record.

B.2 Simulation Exercise

Duration: ~3h

Venue: Technical University of Munich, Institute of Flight System Dynamics, Boltzmannstr. 15, 85748 Garching

We will ask pilots to fly various instrument landing system approaches (CAT I) with changing external parameters regarding air traffic control specifications and environmental effects (e.g. speed constraints, wind). Relevant approach charts are provided in Aeronautical Information Publication. The digital assistant will be available in half of the scenarios. Whether the assistant will be available in the first or second half will be determined by chance.

We will conduct a short questionnaire after each approach. Furthermore, in a post-session discussion, we want to discuss the overall usability and the HMI concept of the digital assistant. We would also like to compare the pilot's self-assessment of the approach with the machine learning model's assessment and a classical flight data monitoring assessment of the approach stability to understand if the machine learning model's assessment correlates with the human situational assessment.

B.3 Research Simulator

The simulator is based on the Dornier DO 728/928 jet-powered regional airliner developed by Fairchild-Dornier. The flight model resembles the performance characteristics of the DO 728, a monoplane design with fixed wings in low wing configuration and two engines mounted under the wings.

The research simulator is a fixed-based design with a high-fidelity visual system and terrain database. The cockpit features two side sticks, rudder paddles, two thrust levers, a gear, a flap, and a speed brake lever. The simulator provides common autopilot and auto-thrust modes, which are selectable on a flight control unit (FCU). However, many systems known from contemporary airliners are not installed, such as Flight Management and Guidance System (FMGS), Traffic Alert and Collision Avoidance System (TCAS), and predictive wind shear system (PWS). The overhead panel consists of a touch screen used for the simulator operation.

During the simulator exercise, a researcher is available on the other pilot seat to assist the pilot with configuration changes, FCU settings, etc. The basic aircraft performance parameters and weather information are provided in Pre-flight Information. As there are no standard operating procedures defined for the research simulator, pilots are requested to act according to their operator policies and procedures.
However, as there exist inconsistent, unstable approach definitions across airlines, we define a stabilization gate at 1000ft AAL for each approach to enable comparability of approaches.

Figure 29: Research Simulator Cockpit

B.4 Data we will record

Non-optional: We collect data in four ways: questionnaires, flight data recorded in the simulator, protocols/written minutes, and post-session interview audio recordings.

Optional: If you consent separately to the non-optional data recording, we would like to video record (incl. audio) the simulation exercises so that the recorded material can be re-consulted during the evaluation of the simulator exercises. The records serve as a redundancy in case we miss documenting something in the minutes or protocols. The videos will only be used to complement the minutes in text form. The **videos will not be published** and will be **deleted after the evaluation of the simulator exercise** and **no later than 30.06.2025** (project end).

The following provides **examples** for each of the non-optional data collection methods:

- an approach self-assessment: In a questionnaire, we will ask you to self-assess the stability of an approach, including possible factors for instability.

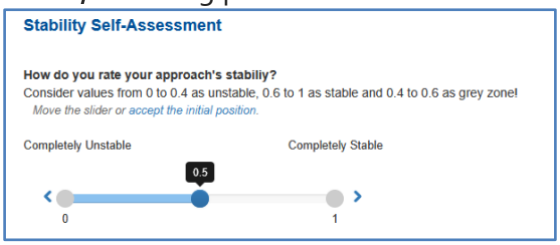


Figure 30: Example Question in Questionnaire

- relevant parameters determining the stability of an approach from the simulator, e.g. Speeds, Configuration, ILS deviations, Cockpit Inputs, …
- During the exercises, we will write minutes for later analysis. E.g. in the summarizing

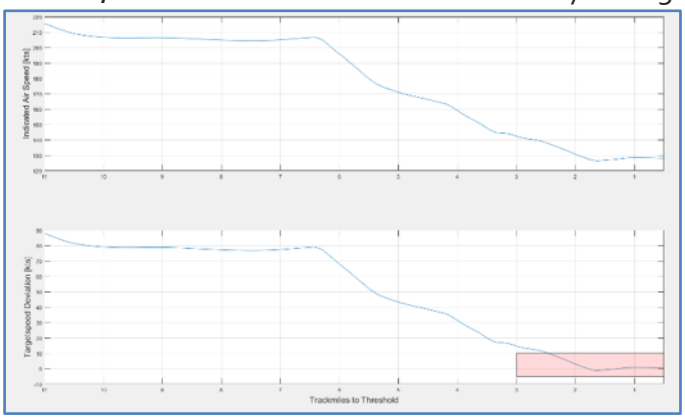


Figure 31: Example of Recorded Data in the Simulator, Indicated Airspeed and Target Speed Deviation, depending on the Distance to Threshold with stabilization criteria as red box.

interview, we aim to discuss the usability of the digital assistant based on your perception during the simulator exercises.

-

B.5 Anonymization of Data

We need to record personal data in the consent form, which contains your name. However, no personal data will be recorded in the questionnaires, minutes and simulator data. We will

not link your personal data from the consent form to any recorded data listed in the section above. Therefore, we ask you not to state the date of the event in the consent form.
The interview audio recordings will be transcribed and fully anonymized.

B.6 Use of recorded Data

We plan to use the recorded data to generate results of the following:

- Graphically visualize approach properties (see e.g. Figure 31),
- Statistically analyze the data (as recorded in the questionnaires, e.g. Figure 30),
- Qualitatively evaluate the written minutes and questionnaire results.
- Compile and evaluate the interview transcriptions.

The results of the simulation exercises will be part of a deliverable report to the European Commission and potential scientific publications.
After a review from the EU, the report will be published on the CORDIS – EU Research Results website https://cordis.europa.eu/project/id/101069877/results and potentially published conference proceedings or journal papers. Excerpts from the interview transcriptions may be included in anonymized form.

B.7 Non-Optional Consent

- I have been told why the study is being done, how information is collected and used.
- I have been allowed to ask questions about the study before it takes place, and know who to contact if I have further questions.
- I know that my participation is voluntary and anonymous and that I can cancel my participation at any time.
- I agree to participate in the study.

Place (no date):_____

Signature:_____

Name of participant in block letters:_____

B.8 Optional Consent for Video Recordings

- I further agree that the simulation exercise is video (incl. audio) recorded only for the purposes stated in the section: Data we will record

Signature:_____

Name of participant in block letters:_____

B.9 Pre-flight Information

Metar EDDM 080/08 030V330 9999 FEW026 SCT035 BKN048 12/04 Q1020

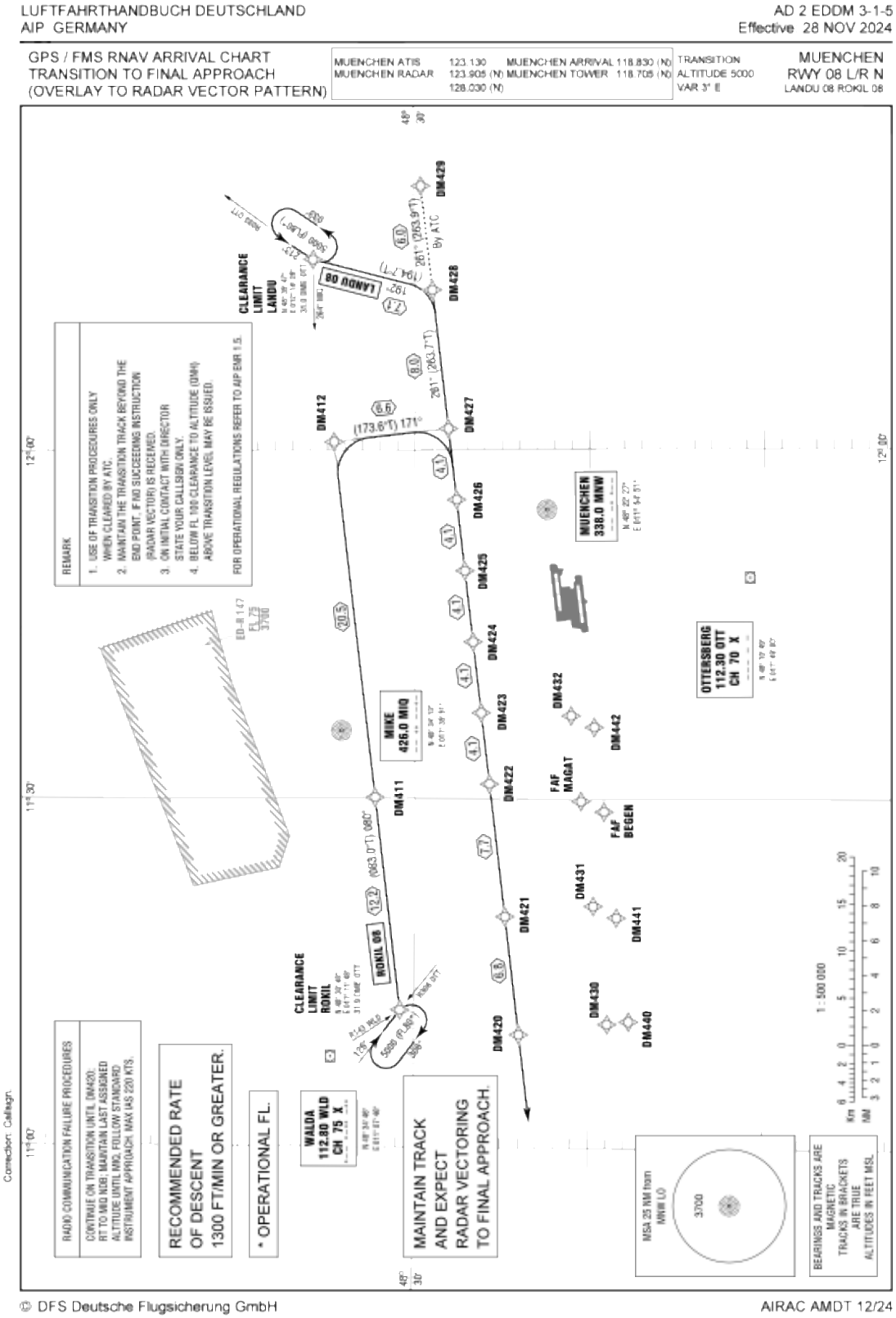
This table comprises a guideline to the basic aircraft performance parameters (pitch/power values vs. aircraft configuration):

Speed (kts)	Pitch (°)	Power N1 (%)	Config
Reference: 5000 ft, 12 NM final			
250	2,0	64	clean
210	3,5	60	clean
170	3,5	62	Flaps 1
170	1,5	61	Flaps 2
Reference: 3° GS			
160	-1,5	34	Flaps 2
V_{app} 127	-0,5	48	Flaps full Gear down

Config	Flaps 1	Flaps 2	Flaps 3	Flaps full
V_{FE}	217	183	163	151

B.10 Aeronautical Information Publications

From: https://aip.dfs.de/BasicIFR/pages/C004BD.html

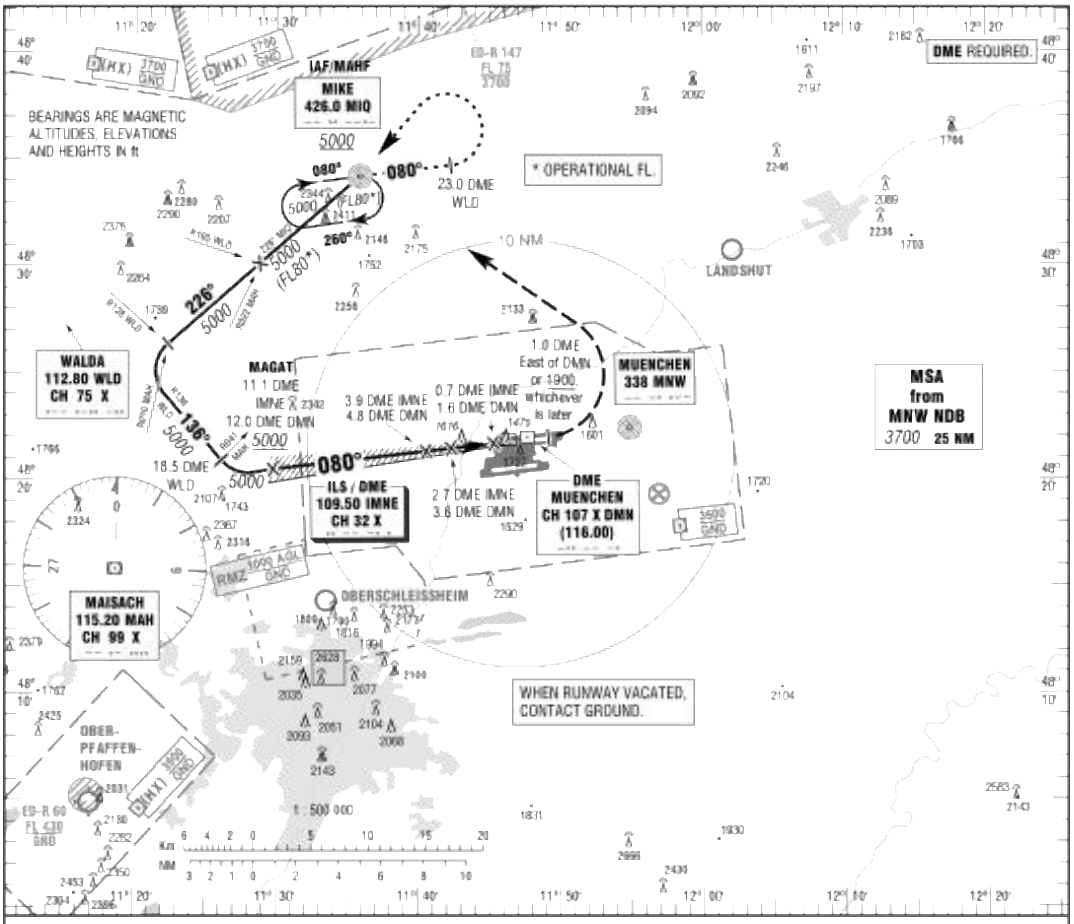


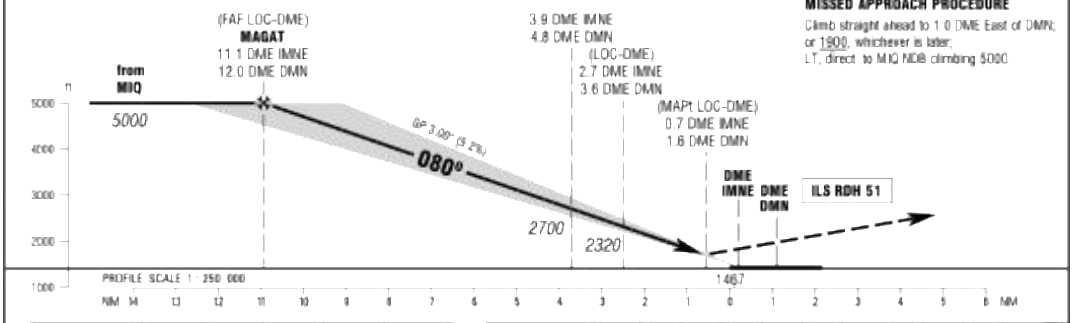
INSTRUMENT APPROACH CHART - ICAO

VAR 3° E | ELEV 1487 | OCH RELATED TO THR 08L ELEV 1467

MUENCHEN ATIS	123.130	MUENCHEN GROUND	121.980 (N)
MUENCHEN RADAR	123.905 (N)	MUENCHEN APRON	121.710
MUENCHEN ARRIVAL	118.830 (N)		121.780
MUENCHEN TOWER	118.705 (N)		121.930

MUENCHEN
ILS CAT II & III or LOC
RWY 08L





OCA (DCH)	A	B	C	D	D_L
ILS/DME CAT I	1606 (139)	1616 (149)	1626 (159)	1636 (169)	
ILS/DME CAT II	1518 (51)	1534 (67)	1547 (80)	1560 (93)	
LOC-DME	1870 (400)	1870 (400)	1870 (400)	1870 (400)	

CAT III APPROVED (MNM RVR 75M).

DME IMNE	11	10	9	8	7	6	5	4	3	2	1
DtST THR	10.8	9.8	8.8	7.8	6.8	5.8	4.8	3.8	2.8	1.8	0.8
ALTITUDE	4960	4640	4320	4010	3690	3370	3050	2730	2410	2100	1780

Timing not authorized for defining the MAPt.

GS	kt	80	100	120	140	160	180
11.1 DME IMNE - THR (10.9 NM)	MIN:SEC	8:11	6:32	5:27	4:40	4:05	3:38
Rate of descent (5.2%)	ft / MIN	420	530	640	740	850	960

Correction: Callsign, obstacles revised.

© DFS Deutsche Flugsicherung GmbH

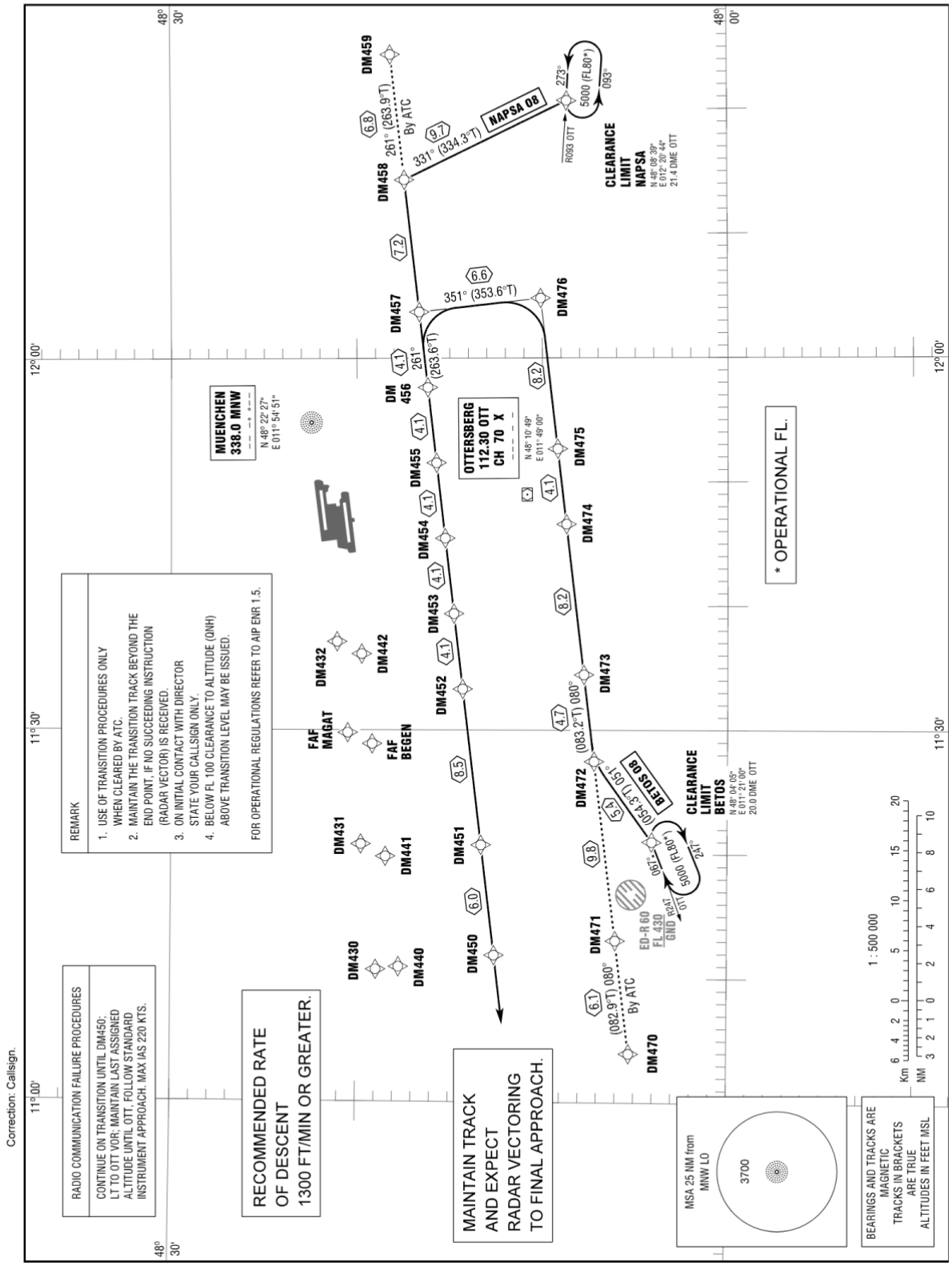
AIRAC AMDT 12/24

GPS / FMS RNAV ARRIVAL CHART
TRANSITION TO FINAL APPROACH
(OVERLAY TO RADAR VECTOR PATTERN)

TRANSITION
ALTITUDE 5000
VAR 3° E

MUENCHEN ATIS	123.130	MUENCHEN ARRIVAL	132.305 (S)
MUENCHEN RADAR	127.955 (S) 120.780 (S)	MUENCHEN TOWER	120.505 (S)

MUENCHEN
RWY 08 L/R S
BETOS 08 NAPSA 08



© DFS Deutsche Flugsicherung GmbH

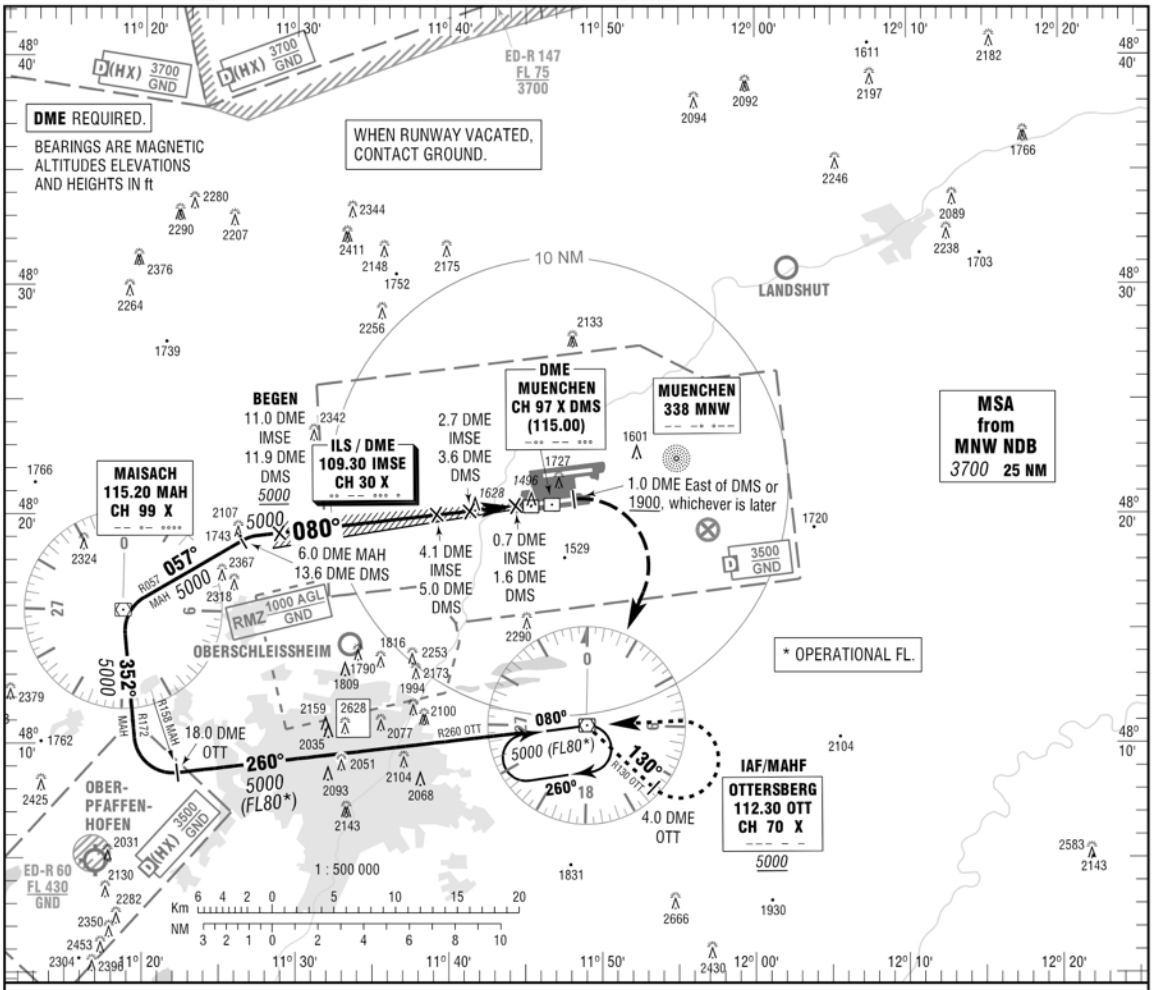
AIRAC AMDT 12/24

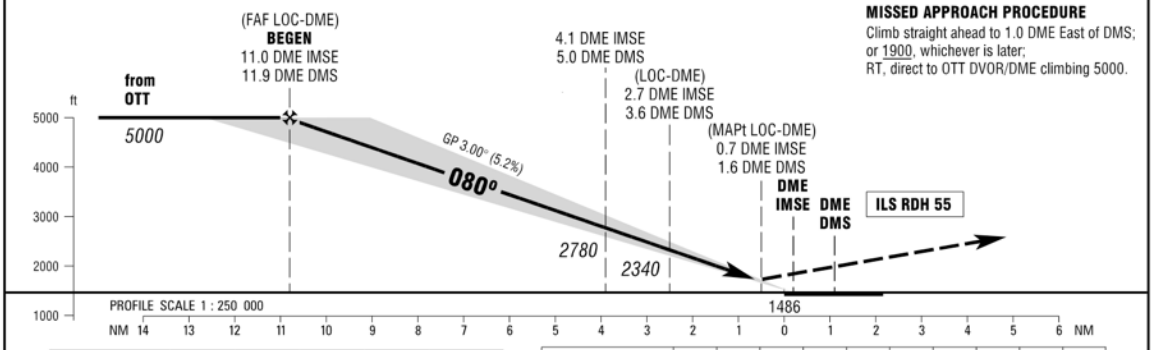
INSTRUMENT APPROACH CHART - ICAO

MUENCHEN ATIS	123.130	MUENCHEN GROUND	121.830 (S)
MUENCHEN RADAR	127.955 (S)	MUENCHEN APRON	121.710
MUENCHEN ARRIVAL	132.305 (S)		121.780
MUENCHEN TOWER	120.505 (S)		121.930

ELEV 1487
OCH RELATED TO THR 08R ELEV 1486
VAR 3° E

MUENCHEN
ILS CAT II & III or LOC
RWY 08R





MISSED APPROACH PROCEDURE
Climb straight ahead to 1.0 DME East of DMS; or 1900, whichever is later;
RT, direct to OTT DVOR/DME climbing 5000.

OCA (OCH)	A	B	C	D	D_L
ILS/DME CAT I	1628 (142)	1638 (152)	1648 (162)	1657 (171)	
ILS/DME CAT II	1539 (53)	1556 (70)	1569 (83)	1582 (96)	
LOC-DME	1880 (390)	1880 (390)	1880 (390)	1880 (390)	

CAT III APPROVED (MNM RVR 75M).

DME IMSE	10	9	8	7	6	5	4	3	2	1
DIST THR	9.8	8.8	7.8	6.8	5.8	4.8	3.8	2.8	1.8	0.8
ALTITUDE	4660	4340	4020	3710	3390	3070	2750	2430	2110	1800

Timing not authorized for defining the MAPt.

GS	kt	80	100	120	140	160	180
11.0 DME IMSE - THR (10.9 NM)	MIN:SEC	8:11	6:32	5:27	4:40	4:05	3:38
Rate of descent (5.2%)	ft / MIN	420	530	640	740	850	960

Correction: Callsign, obstacles revised.

© DFS Deutsche Flugsicherung GmbH

AIRAC AMDT 12/24

Appendix C Data Analysis Materials

C.1 Thematic Analysis LLM Coding Prompt

"You are an expert qualitative researcher with extensive experience in thematic analysis. Please generate as many initial inductive codes as needed from the provided interview transcript, capturing semantic and latent meanings. Code only the responses from the participant ("pilot"). Ignore the interviewer's ("researcher") comments. Each code should include: a short name, a 2-3 sentence description, and 1-3 supporting quotes (max 40 words each). Only use quotes specifically by the participant, not the interviewer-researcher. Don't rearrange quotes, but shorten them if needed (using "..." to indicate removed sections) to maintain quote clarity and conciseness.
Output the results in an Excel spreadsheet with columns for Code ID #, Code name, Description, and Quote (with quotation marks). In cases of multiple relevant quotes for a code, separate quotes with commas."

C.2 Thematic Analysis LLM Code Reconciliation Prompt

"I've uploaded:
- An Excel file with LLM-generated codes, descriptions, and quotes from multiple interview transcripts (indicated by participant ID #).
- One or more .pdf transcript files corresponding to participant interviews with my own margin comments (initial impressions or interpretive notes). Each LLM-selected quote in the Excel file is linked to a participant quote in the corresponding transcript.
Please:
Match each LLM-identified code, description, or participant quote in the Excel sheet to my margin comments from the .pdf document(s). For semantically matching items, add my verbatim margin comment/annotation to the column "Researcher Notes". For each item, fill out the column called "Agreement? (Y/N/Partial)" based on semantic alignment between the LLM code/description/quote and my margin comment.
If I've commented in the margin on a segment not covered by any LLM code, insert it as a new row with:
- Code: a short code summarizing the comment (use my margin comment content and the quoted section to formulate an appropriate/relevant code name in the same style as the LLM-generated code names)
- Quote: the quoted transcript segment (participant comment – not my margin comment, nor the interviewer-researcher's comments) (max 50 words)
- Description: 1–2 sentence elaboration (again, use my margin comment content as basis for elaborating a description)
- Analyst: "Human" (otherwise "LLM")
- Researcher Notes: my verbatim margin comment.
Maintain original columns like Code, and Description for structure. Output the results as a structured Excel sheet with both LLM and human insights combined."

C.3 Thematic Analysis LLM Theme Development Prompt

"You are an expert qualitative researcher with extensive knowledge about thematic analysis. I have provided a spreadsheet with thematic codes (ID #, Code name, Code description, representative participant quotes, etc.). Please cluster these codes into broader themes, following best practices for thematic analysis (regarding number of themes, subthemes, groundedness in data, uniqueness, etc.). For each theme, give a name, a 2-5 sentence description, and the codes it includes. Format possible subthemes in the same way and indicate the theme hierarchy. Output the results as an Excel file."