



# Recommendations recovery measures and HMI Implementation (Initial version)

Sara Silvagni, Carlo Valbonesi (DBL), Barry Kirwan (ECTL), Jean-Marc Andre (IPB/ENSC), Jaime Diaz Pineda (CATIE), Alia Lemkaddem, Olivier Grossenbacher (CSEM), Rebecca Charles, Jim Nixon (Cranfield), Frederik Mohrmann (NLR), Nicolas Maille (ONERA), Matthias Wies (DLR)

Future Sky Safety is a Joint Research Programme (JRP) on Safety, initiated by EREA, the association of European Research Establishments in Aeronautics. The Programme contains two streams of activities: 1) coordination of the safety research programmes of the EREA institutes and 2) collaborative research projects on European safety priorities.

This deliverable is produced by the Project P6 "Human Performance Envelope". The main objective is the definition of method and principles to recover performance if outside of the human performance envelope, and provides initial guidelines for innovative Human Machine Interface (HMI) development, taking into account one dedicated concept of automation.

<b>Programme Manager</b>	Michel Piers, NLR
<b>Operations Manager</b>	Lennaert Speijker, NLR
<b>Project Manager (P6)</b>	Matthias Wies, DLR
<b>Grant Agreement No.</b>	640597
<b>Document Identification</b>	D6.4
<b>Status</b>	Approved
<b>Version</b>	2.0
<b>Classification</b>	Public

*This page is intentionally left blank*

## Contributing partners

Company	Name
DBL	Sara Silvagni, Carlo Valbonesi
ECTL	Barry Kirwan
DLR	Matthias Wies
NLR	Frederik Mohrmann, Mick Vermaat, Tanja Bos
ONERA	Nicolas Maille
IPB/ENSC	Jean-Marc Andre
CATIE	Jaime Diaz Pineda
CSEM	Alia Lemkaddem, Olivier Grossenbacher
CRANFIELD UNIVERSITY	Rebecca Charles, Jim Nixon
TAV	Sylvain Hourlier

## Document Change Log

Version	Issue Date	Remarks
1.0	19-05-2017	First formal release
2.0	28-07-2017	Second formal release

## Approval status

Prepared by: (name)	Company	Role	Date
Sara Silvagni	DBL	Main Author	19-05-2017
Carlo Valbonesi	DBL	Main Author	19-05-2017
Checked by: (name)	Company	Role	Date
Alex Rutten	NLR	Quality Assurance	28-07-2017
Approved by: (name)	Company	Role	Date
Matthias Wies	DLR	Project Manager (P6)	19-05-2017
Lennaert Speijker	NLR	Operations Manager	28-07-2017

## Acronyms

Acronym	Definition
AOI	Area of Interest
AP	Application of Procedures
ATC	Air Traffic Controller
ATIS	Automatic Terminal Information Service
ATM	Air Traffic Management
CPT	Captain
CWC	Cross Wind Component
DM	Decision Making
EASp	European Aviation Safety plan
EC	European Council
ECAM	Electronic Centralised Aircraft Monitor
ED	Engine Display
EFB	Electronic Flight Bag
EFIS	Electronic Flight Instrument System
EREA	association of European Research Establishments in Aeronautics
EU	European Union
FAF	Final Approach Fix
FCU	Flight Control Unit
FO	First Officer
FSS	Future Sky Safety
HMI	Human-Machine Interface
HPE	Human Performance Envelope
HR	Heart Rate
HRV	Heart Rate Variability
HUD	Head Up Display
IAF	Intermediate Approach Fix
ICC	Inter Class Correlations
MCDU	Multifunction Control Display Unit
MERIA	MEntal Representation Impact Analysis



<b>MR</b>	Mental Representation
<b>MW LGTS</b>	Master Warning Lights
<b>ND</b>	Navigation Display
<b>OVHD</b>	Overhead panel
<b>PF</b>	Pilot Flying
<b>PFD</b>	Primary Flight Display
<b>PRED</b>	Predicted Performance
<b>PM</b>	Pilot Monitoring
<b>RMP</b>	Radio Management Panel
<b>SA</b>	Situation Awareness
<b>SD</b>	Standard Deviation
<b>SDNN</b>	Standard Deviation of NN intervals
<b>SE</b>	Standard Error
<b>SESAR</b>	Single European Sky ATM Research
<b>SRIA</b>	Strategic Research and Innovation Agenda
<b>TOD</b>	Top of Descent
<b>TRA</b>	Temporal Reliability Analysis
<b>WL</b>	Workload
<b>WSHLD LFT / RGT</b>	Windshield left / right

## EXECUTIVE SUMMARY

### Problem Area

Performance degradation in the cockpit is still an important safety issue, particularly during flight 'upset' conditions, which is one of the top safety risks in commercial aviation. In the past, for at least a couple of decades following automation and the introduction of the 'glass cockpit', the main safety concern with pilots was with crew resource management, i.e. ensuring that the flight crew worked together as an effective team. Today cockpits are increasingly automated, including not only the cockpit dashboard but also electronic flight bags. Two major areas of concern are prevalent today – how do pilots maintain shared situation awareness given a highly automated flight deck, and how do we ensure they react safely when a flight upset occurs? The HPE project merges these two concerns by exploring how pilots maintain effective and safe shared situation awareness during flight upset conditions. The intended 'endgame' of the project is to be able to protect the 'human performance envelope' prior to and during such events, so that their performance does not degrade when we need them most, and to ensure that automation is optimally supportive in helping the pilots maintain shared situation awareness and hence return the aircraft to a safe state.

The Human Performance Envelope (HPE) concept is based on the assumption that people's performance is shaped by the influence of a set of interdependent factors (e.g. workload, stress, situation awareness, fatigue, etc.). According to this concept, a small variation in some of these interdependent factors may generate a greater influence on operator's performance than a big variation of one single factor in isolation. If these factors, working alone or in combination, are studied borrowing the envelope metaphor, it can be possible to determine the starting point in which significant performance degradation could affect safety.

The HPE concept is a novel and interesting approach to safe human performance, and although it has intuitive appeal, research is needed to determine if there is sufficient scientific evidence that the HPE is valid, that it can be used to measure and/or predict when pilot performance degrades, and whether it can help inform automation strategies for future cockpit design. These can be expressed as questions or formal hypotheses:

1. Is the HPE concept valid?
2. Can it be used to detect and/or predict pilot performance degradation?
3. Can it inform flight deck automation design strategies?

To test the HPE concept, a real-time simulation with 10 First Officers from a major European airline was conducted at a DLR research full-scope, moving flight simulator in May, 2016. The simulation was split in two parts. The first part was focused on providing data to validate the HPE concept, and consisted of short 'runs' (eight runs for each pilot, of around 20 minutes) where three factors were progressively increased (workload, stress and reduced situation awareness), and the pilots had to deal with these challenges. The second part involved a longer scenario (e.g. 40 minutes) where more happened as 'mini-episodes' during

the run. The first part, called Scenario 1 (with eight runs) was designed to test the HPE concept and see which measures worked best at detecting performance degradation. The second Scenario was aimed more at exploring how the flight deck HMI supported (or did not support) the resolution of flight upset conditions.

A first round of analysis for behavioural, psycho-physiological, performance-based and subjective data was performed to determine points where human performance deteriorates, as well as to identify behavioural and/or physiological markers critical in signalling performance degradation. This first analysis was essential, since the three questions above cannot be answered if the simulation was not eliciting pilot performance degradation. In fact the simulation worked well, and was able to 'push' the pilots toward the edges of their performance envelope, and certainly the upset conditions simulated pushed many of the pilots out of their 'comfort zone.' The results from these analyses are reported in D6.3 "Test report preliminary testing with system pilots' cognitive task analysis".

The simulation runs therefore produced a vast array of data – performance, behavioural and physiological – and this deliverable aims at advancing the first round of analysis through the triangulation of the different datasets in order to:

- Validate the HPE concept and test its applicability to different operators and tasks;
- Identify reliable measures able to measure and (if possible) predict performance variation, to be employed in the final P6 simulation;
- Use the data collected and the HPE concept to define potential improvements in the HMI able to support recovery from performance degradation.

## Description of Work

This report completes the analysis of data collected during the first FSS P6 simulations, held at DLR research simulator in May 2016 with 10 First Officers from a major European airline. Specifically, the different datasets from the simulation runs are correlated to:

- Prove the HPE model in a partially controlled simulation setting (Scenario 1), via:
  - Correlation between runs and three HPE factors (e.g. workload, measured through subjective ratings), runs and physiological factors (e.g. heart rate), and runs and objective performance (e.g. deviation from glideslope and localiser) and;
  - Correlation between performance and physiological factors to identify a potential equation able to predict the performance through the analysis of pilot's status.
- Test the HPE model in an ecologically valid setting, basically taking the abovementioned predictive equation based on Scenario 1 and trying to apply it to Scenario 2, using a different task and different performance measures.
- Use the results from the previous sections to identify performance decrement areas and improve HMI to support pilot performance recovery.

## Results & Conclusions

The different analyses performed showed that there are links between the three components which are supposed to shape the human envelope, as all the factors vary in each run of Scenario 1. The increase of workload is always associated to an increase of the stress level and a decrease of the situation awareness. The three factors were manipulated alone or in combination, and results from the combination runs produced a greater effect on performance than the single factor manipulation, despite the fact that in the single factor runs the other factors are indirectly affected as well. This provided evidences that the HPE concept works, as a small variation in several factors at the same time (e.g. medium level of stress, medium level of workload and medium level of SA) pushed the envelope and provoked a greater performance decrease than a big variation on a single factor (e.g. very high level of workload).

Uncertain results emerged from the analysis of physiological data. In fact, while the simulation confirmed a correlation between heart rate and pupil diameter with Workload and Stress, so linear models can be created between the two variables, results on heart rate variability were conflicting, and variation on situation awareness could not be detected using physiological signals. Eye tracking data gave better results with this respect, as the analysis of scan path and heat maps can give us useful information on Situation Awareness degradation and how the interface is used. However, the correlation tasks showed that physiological data cannot be used alone to recognise or to predict performance decrement, as the performance measures are task dependent and subjective variability play a big role on these data.

## Applicability

Despite the lack of predictability, physiological data presented some interesting features and can be associated to factors' variation. Thus their use in the next simulations therefore should not be discounted. However, attention has to be paid on how the data are used (normalised per pilot, for factors validation instead of performance prediction) and the conclusions that can be derived from them, especially on pilot situation awareness.

The HMI issues connected to performance decrements will be addressed by Work Package 6.4, specifically dedicated to the development of ways to augment the envelope and to the design of future cockpit concepts.

Overall, the analysis so far has provided partial evidence for the HPE concept and the detection (but not predictability) of performance degradation. Following the final simulation in P6 at the end of 2017, the report D6.4b will be able to be more conclusive on the HPE's validity, its transferability from one scenario to another, and its utility for safeguarding human performance in flight upset conditions.

*This page is intentionally left blank*

## TABLE OF CONTENTS

Contributing partners	3
Document Change Log	3
Approval status	3
Acronyms	4
<b>Executive Summary</b>	<b>6</b>
Problem Area	6
Description of Work	7
Results & Conclusions	8
Applicability	8
<b>List of Figures</b>	<b>13</b>
<b>List of Tables</b>	<b>16</b>
<b>Introduction</b>	<b>17</b>
1.1. The Programme	17
1.2. Project context	17
1.3. Research objectives	18
1.4. Approach	19
1.5. Structure of the document	19
<b>2 Factors effect on the Envelope</b>	<b>22</b>
2.1. Global effects of Workload, Stress and Situation Awareness on the envelope	22
2.1.1. Global effects of workload	25
2.1.2. Global effects of stress	30
2.1.3. Global effects of degraded situation awareness	33
2.1.4. Conclusions about the HP evolution	37
2.2. Global effects of combined increase of workload, stress and degraded situation awareness	39
2.2.1. Conclusions about the global effect of combined increase of workload, stress and degraded situation awareness	41
2.3. Short term effects of Stress and Situation Awareness on the envelope	42
2.3.1. Short term effect of Stress	42
2.3.2. Short term effect of reduced situation awareness	44
2.3.3. Conclusions about short term effects	45
2.4. Situation Awareness analysis using scan path and eye tracking data	46
2.4.1. Use of eye tracking for HCI research and design	47
2.4.1.1. Scan Pattern	47
2.4.1.2. Fixation duration	47

2.4.1.3.	Number of fixations	48
2.4.2.	Overview of tasks in Run 8	48
2.4.3.	Performance characteristics of the ideal timeline	49
2.4.4.	Method	51
2.4.4.1.	Participants	51
2.4.4.2.	Eye Tracking Technology	51
2.4.4.3.	Approach to Analysis	52
2.4.5.	Pilot timelines	53
2.4.6.	Differences between pilots for whole run	55
2.4.7.	Differences between pilots for specific events	55
2.4.7.1.	Delay Vectors	56
2.4.7.2.	Localiser Interference	56
2.4.7.3.	Loud Noise	56
2.4.7.4.	Wind shift	56
2.4.8.	Pilot deep-dives	56
2.4.8.1.	Key events and dialogue - Pilot 6	57
2.4.9.	Conclusions	57
<b>3</b>	<b>HPE Model Validation</b>	<b>59</b>
3.1.	Correlating HPE and performance	60
3.2.	Correlating HPE and physiological data	62
3.2.1.	Workload as measured by ISA	63
3.2.2.	Workload as measured by NASA-TLX	64
3.2.3.	Stress measured by SACL	65
3.2.4.	Situation Awareness measured by SART	66
3.2.5.	Conclusions	66
3.3.	Correlating performance and physiological data	67
3.4.	Validating HPE concept with Scenario 2	69
3.4.1.	From physiological data to predicted performance	69
3.4.2.	Comparing predicted and actual performance	71
3.4.3.	Conclusions on the applicability of the HPE model	74
3.4.4.	From actual performance to physiological data	74
<b>4</b>	<b>Principles and Considerations for HMI Design to support recovery from Performance Degradation</b>	<b>75</b>
4.1.	Support the recovery of Pilot Flying performance	75
4.2.	Support the recovery of Pilot Monitoring performance	77
4.2.1.	Pilots Mental Representations	77
4.2.1.1.	Fuel status	82
4.2.1.2.	Electrical failure	82

4.2.1.3.	Weather	84
4.2.2.	Competence evaluation	85
4.2.3.	Results from competence evaluation and cognitive walkthrough matching	90
4.2.4.	Considerations for HMI design in support of Pilot Monitoring	95
<b>5</b>	<b>Conclusions and recommendations</b>	<b>97</b>
5.1.	Conclusions	97
5.2.	Recommendations	101
<b>6</b>	<b>References</b>	<b>103</b>
<b>Appendix A</b>		
	<b>Situatuion Awareness and Eye Tracking Data</b>	<b>104</b>
<b>Appendix A.1</b>	Stacked Bar Charts for Key Events	104
<b>Appendix A.2</b>	Flight deck dialogue Pilot 6	106
<b>Appendix A.3</b>	Detailed deep dive and heat-map analysis	109



## LIST OF FIGURES

FIGURE 1: STRUCTURE OF THE DOCUMENT .....	20
FIGURE 2: OVERVIEW OF THE RESEARCH PERSPECTIVES ADOPTED IN THIS DOCUMENT FOR DATA COLLECTION AND ANALYSIS AND EXTERNAL STAKEHOLDERS .....	21
FIGURE 3: NASA-TLX AND ISA MEASURES OF THE WORKLOAD .....	25
FIGURE 4: 10D-SART SITUATION AWARENESS AND SACL STRESS MEASURES .....	26
FIGURE 5: HR (LEFT) AND SDNN (RIGHT) FOR NORMALISED DATA IN PHASE 2 .....	26
FIGURE 6: NORMALISED PUPIL RADIUS .....	27
FIGURE 7: SELF-ESTIMATED MEDIAN PERFORMANCE .....	27
FIGURE 8: LOCALISER AND GLIDESLOPE DEVIATIONS .....	28
FIGURE 9: PERCENTAGE OF GO-AROUND BY RUNS .....	28
FIGURE 10: MODIFICATION OF THE ENVELOPE .....	29
FIGURE 11: EVOLUTION OF THE STRESS LEVEL .....	30
FIGURE 12: WORKLOAD .....	30
FIGURE 13: SITUATION AWARENESS .....	31
FIGURE 14: NORMALISED HR AND SDNN .....	31
FIGURE 15: NORMALISED MEAN EYE RADIUS .....	31
FIGURE 16: SELF-ESTIMATED MEDIAN PERFORMANCE .....	32
FIGURE 17: LOCALISER AND GLIDE-SLOPE DEVIATIONS .....	32
FIGURE 18: MODIFICATION OF THE ENVELOPE .....	33
FIGURE 19: SART SITUATION AWARENESS .....	34
FIGURE 20: EVOLUTION OF THE WORKLOAD .....	34
FIGURE 21: STRESS LEVEL .....	34
FIGURE 22: HEART RATE AND HEART RATE VARIATIONS .....	35
FIGURE 23: EYE RADIUS .....	35
FIGURE 24: SELF-ESTIMATED PERFORMANCES .....	36
FIGURE 25: LOCALISER AND GLIDE-SLOPE DEVIATIONS .....	36
FIGURE 26: GO-AROUND .....	36
FIGURE 27: MODIFICATION OF THE ENVELOPE .....	37
FIGURE 28: WL, STRESS AND SA MODIFICATION AND RESULTING HP ENVELOPE EVOLUTIONS .....	38
FIGURE 29: WORKLOAD .....	39
FIGURE 30: SITUATION AWARENESS AND STRESS .....	39
FIGURE 31: HEART RATE AND HEART RATE VARIABILITY .....	39
FIGURE 32: NORMALISED MEAN EYE RADIUS .....	40
FIGURE 33: SELF-ESTIMATED PERFORMANCES .....	40
FIGURE 34: LOCALISER AND GLIDE-SLOPE DEVIATIONS .....	41
FIGURE 35: PERCENTAGE OF GO-AROUND .....	41
FIGURE 36: EVOLUTION OF THE HP ENVELOPE .....	42

FIGURE 37: PUPIL RADIUS AND AREAS OF INTEREST AROUND THE BEGINNING OF THE LOUD NOISE (RUN 8, PILOT 6).....	43
FIGURE 38: INCREASE OF THE PUPIL RADIUS AFTER THE BEGINNING OF THE LOUD NOISE FOR EACH PILOT AND EACH RUN WHERE A NOISE WAS INTRODUCED .....	43
FIGURE 39: INCREASE OF THE HEART RATE AFTER THE BEGINNING OF THE LOUD NOISE FOR EACH PILOT AND EACH RUN WHERE A NOISE WAS INTRODUCED .....	44
FIGURE 40: TYPICAL EVOLUTION OF THE HEART RATE DURING THE LANDING FLIGHT PHASE .....	44
FIGURE 41: INCREASE OF THE PUPIL RADIUS AFTER THE BEGINNING OF LOCALISER INTERFERENCES FOR EACH PILOT AND EACH RUN WHERE LOCALISER INTERFERENCES WERE INTRODUCED .....	45
FIGURE 42: MODEL OF SITUATION AWARENESS (ENDSLEY, 1995) .....	46
FIGURE 43: AREAS OF INTEREST .....	49
FIGURE 44: AOIs GROUPED BY CONTROL FUNCTION .....	51
FIGURE 45: DATA PROCESSING PROCEDURE FOR TIMELINE GENERATION.....	52
FIGURE 46: DATA PROCESSING PROCEDURE FOR DEEP DIVE ANALYSIS .....	53
FIGURE 47: TIMELINE FOR PILOT 6.....	54
FIGURE 48: TIMELINE FOR PILOT 8.....	54
FIGURE 49: DWELL TIMES PER AOI FOR WHOLE OF RUN 8 .....	55
FIGURE 50: EXAMPLE OF HEAT-MAP (AOI FREQUENCY AND DIRECTION PILOT 6 FROM 4:32 TO 6:03 MINUTES) .....	57
FIGURE 51: CORRELATION TASKS TO VALIDATE HPE MODEL AND PREDICT PERFORMANCE IN SCENARIO 2 .....	60
FIGURE 52: CORRELATION BETWEEN PERFORMANCE AND PHYSIOLOGICAL DATA .....	68
FIGURE 53: PILOT 5 HEART RATE (HR), PUPIL DIAMETER (ET) AND PREDICTED PERFORMANCE (PRED) .....	70
FIGURE 54: PILOT 10 HEART RATE (HR), PUPIL DIAMETER (ET) AND PREDICTED PERFORMANCE (PRED).....	70
FIGURE 55: PILOT 5 PERFORMANCE CORRELATION ANALYSIS (PRED-SA) .....	71
FIGURE 56: PILOT 5 PERFORMANCE CORRELATION ANALYSIS (PRED-DM) .....	72
FIGURE 57: PILOT 5 PERFORMANCE CORRELATION ANALYSIS (PRED-AP).....	72
FIGURE 58: PILOT 10 PERFORMANCE CORRELATION ANALYSIS (PRED-SA) .....	73
FIGURE 59: PILOT 10 PERFORMANCE CORRELATION ANALYSIS (PRED-DM) .....	73
FIGURE 60: PILOT 10 PERFORMANCE CORRELATION ANALYSIS (PRED-AP).....	73
FIGURE 61: STRUCTURE OF SCENARIO 2 AND MENTAL REPRESENTATION OF THE PILOT .....	77
FIGURE 62: NODES COLUMN - OMB PROCEDURE NOT PERFORMED .....	78
FIGURE 63: MENTAL REPRESENTATION OF PILOT 5 .....	79
FIGURE 64: EMERGENCY IS DECLARED LATE .....	80
FIGURE 65: WARNING MESSAGE OF LAPA RWY09 .....	81
FIGURE 66: INPUT - LAPA RESULTS AND CROSSWIND LIMITATIONS.....	81
FIGURE 67: FUEL STATE IN THE HMI .....	82
FIGURE 68: ELECTRICAL FAILURE SITUATION .....	83
FIGURE 69: ECAM STATUS OF BUS FAILURE.....	83
FIGURE 70: CONSIDERATIONS FOR LAPAs COMING FROM ELECTRICAL FAILURE .....	84
FIGURE 71: TWO DIFFERENT HMI MESSAGES MEANING THAT ROLL-OUT MUST BE IN MANUAL MODE .....	85
FIGURE 72: LIMITATIONS FOR LANDING IN OM-B; CATII PROCEDURE.....	85

FIGURE 73: TRA OF PILOT 5 SITUATIONAL AWARENESS.....	87
FIGURE 74: TRA OF PILOT 5 DECISION MAKING.....	87
FIGURE 75: TRA OF PILOT 5 APPLICATION OF PROCEDURES.....	88
FIGURE 76: TRA OF PILOT 3 DECISION MAKING.....	88
FIGURE 77: TRA OF PILOT 4 DECISION MAKING.....	89
FIGURE 78: TRA OF PILOT 10 DECISION MAKING.....	89
FIGURE 79: OVERVIEW OF MERIA MODEL ITEMS RELATED TO SA OR DM.....	91
FIGURE 80: COMBINING MERIA MODEL SA-ITEMS (COLOUR BARS) WITH NLR'S SA TRA.....	92
FIGURE 81: COMBINING MERIA MODEL DM-ITEMS (COLOUR BARS) WITH NLR'S DM TRA.....	92
FIGURE 82: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 0 TO 2:16 MINUTES.....	110
FIGURE 83: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 2:56 TO 3:36 MINUTES.....	111
FIGURE 84: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 3:37 TO 3:53 MINUTES.....	111
FIGURE 85: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 4:32 TO 6:03 MINUTES.....	112
FIGURE 86: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 6:03 TO 8:01 MINUTES.....	113
FIGURE 87: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 8:55 TO 8:59 MINUTES.....	113
FIGURE 88: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 8:59 TO 9:35 MINUTES.....	114
FIGURE 89: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 9:36 TO 10 MINUTES.....	114
FIGURE 90: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 10:27 TO 10:58 MINUTES.....	115
FIGURE 91: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 12:10 TO 12:34 MINUTES.....	116
FIGURE 92: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 14:34 TO 14:53 MINUTES.....	116
FIGURE 93: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 15:40 TO 15:55 MINUTES.....	117
FIGURE 94: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 19:02 TO 19:12 MINUTES.....	117
FIGURE 95: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 19:41 TO 20:22 MINUTES.....	118
FIGURE 96: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 20:25 TO 21:25 MINUTES.....	119
FIGURE 97: AOI FREQUENCY AND DIRECTION PILOT 6 FROM 23:07 TO 23:31 MINUTES.....	120

## LIST OF TABLES

TABLE 1: SCENARIO 1 RUNS, HPE FACTORS SUPPOSEDLY AFFECTED AND AIRPORT .....	22
TABLE 2: AOI ACRONYM DEFINITIONS.....	50
TABLE 3: REGRESSION TABLE OF HPE AND PERFORMANCE CORRELATION .....	62
TABLE 4: ISA MULTIPLE LINEAR REGRESSION ANALYSIS (N = 32).....	63
TABLE 5: NASA-TLX MULTIPLE LINEAR REGRESSION ANALYSIS (N=32) .....	64
TABLE 6: NASA-TLX MULTIPLE LINEAR REGRESSION ANALYSIS WITHOUT SDNN (N=32).....	65
TABLE 7: SACL MULTIPLE LINEAR REGRESSION ANALYSIS (N=32).....	65
TABLE 8: SART LINEAR REGRESSION ANALYSIS (N=32).....	66
TABLE 9: PILOT 5 PERFORMANCE CORRELATION ANALYSIS.....	71
TABLE 10: PILOT 10 PERFORMANCE CORRELATION ANALYSIS .....	72
TABLE 11: OVERVIEW OF SA PERFORMANCE-CORRECTED MR RATINGS PER FLIGHT PHASE .....	93
TABLE 12: OVERVIEW OF DM PERFORMANCE-CORRECTED MR RATINGS PER FLIGHT PHASE .....	94
TABLE 13: OVERVIEW OF DM PERFORMANCE-CORRECTED MR RATINGS PER FLIGHT PHASE .....	95
TABLE 14: SUMMARY OF CORRELATION TASKS BETWEEN RUNS AND HPE / PHYSIOLOGICAL DATA / PERFORMANCE DATA .....	99

## INTRODUCTION

### 1.1. The Programme

The EC Flight Path 2050 vision aims to achieve the highest levels of safety to ensure that passengers and freight as well as the air transport system and its infrastructure are protected. Trends in safety performance over the last decade indicate that the ACARE Vision 2020 safety goal of an 80% reduction of the accident rate is not being achieved. A stronger focus on safety is required.

Future Sky Safety, established under coordination of EREA, is a Transport Research Programme built on European safety priorities that brings together 33 European partners to develop new tools and new approaches to aeronautics safety. The Programme links the EASp (European Aviation Safety plan) main pillars (operational issues, systemic issues, human performance and emerging issues) to the Flight Path 2050 safety challenges through four Themes:

- **Theme 1** (new solutions for today's accidents) aims for breakthrough research to address the current main accident categories in commercial air transport with the purpose of enabling a direct, specific, significant risk reduction in the medium term.
- **Theme 2** (strengthening the capability to manage risk) conducts research on processes and technologies to enable the aviation system actors to achieve near-total control over the safety risk in the air transport system.
- **Theme 3** (building ultra-resilient systems, organizations and operators) conducts research on the improvement of Systems, Organisations and the Human Operator with the specific aim to improve safety performance under unanticipated circumstances.
- **Theme 4** (building ultra-resilient vehicles) aims at reducing the effect of external hazards on vehicle integrity as well as reducing the number of fatalities in case of accidents.

Together, these Themes and the institutionally funded safety research intend to cover the safety priorities of Flight Path 2050 as well as the ACARE Strategic Research and Innovation Agenda (SRIA) (in particular the Challenges brought forward by ACARE Working Group 4 "Safety and Security").

The Programme will also help coordinate the research and innovation agendas of several countries and institutions, as well as create synergies with other EU initiatives in the field (e.g. SESAR, Clean Sky 2). Future Sky Safety is set up with expected seven years duration, divided into two phases of which the first one of 4 years has been formally approved.

### 1.2. Project context

Future Sky Safety P6 addresses Theme 3 (Building ultra-resilient systems and operators) focussed on strengthening the resilience to deal with current and new risks of the humans and the organizations operating the air transport system.

P6 builds on a concept previously proposed in the Air Traffic Management (ATM) domain, extending it to the Human Operators in the cockpit. The aim of the project is to define and apply the Human Performance Envelope for cockpit operations and design, and determining methods to recover crew's performance to the centre of the envelope, and consequently to augment this envelope.

The Human Performance Envelope is to some extent a new paradigm in Human Factors. Rather than focusing on one or two individual factors (e.g. fatigue, situation awareness, etc.), it considers a range of common factors in accidents and maps how they work alone or in combination to lead to a performance decrement that could affect safety. The safe region on the envelope is bordered by markers, which can be measured and signalled, allowing the pilots to detect and recover, or enabling external agencies to prompt recovery, or allowing automation to kick in and take over. The Human Performance Envelope will deal with the most crucial people in the accident chain, giving them back-up when they most need it, assuring performance when things get difficult. It will increase safety by focusing on the sharp end of accidents, and consign the term 'Pilot error' to the waste paper bin. The impact will primarily be through improved design and operational practices and is thus expected in the short to medium term.

### 1.3. Research objectives

FSS Project P6's main goal is to define and apply the concept of the Human Performance Envelope in the terms of cockpit operations and design. Based on the current knowledge about cognitive demands in the cockpit, the project will determine methods to restore the crew's performance to the centre of the envelope, and consequently to augment this envelope, through innovative HMI design, new automation concepts and new flight crew monitoring solutions (with impact on procedures or training).

In particular, by the end of the Project P6 the following results are expected:

- New Guidelines for HMI development, taking into account one dedicated concept of automation.
- General Guidelines for Augmenting the Envelope.
- Demonstrator (i.e. prototype with limited functionalities in an example scenario) of HPE monitoring and regulation solutions implemented in full mission simulators.

During the first simulations, held at DLR research simulator in May 2016 with 10 First Officers from a major European airline, a large set of behavioural, psycho-physiological, performance-based and subjective data were collected. A preliminary analysis was then conducted on each group of data to compare the different runs and look at the impact that each single factor or the combination of factors had on pilots' status and performance. Results from this analysis can be found in D6.3.

## 1.4. Approach

In this second report, the psycho-physiological, performance-based and subjective data are analysed to look at the effects of each factor on the other factors and on the global human performance, and to identify areas where performance decrements.

Also, the report contains the results of the validation of the HPE model, performed through a set of correlations that explore the relation between HPE, Performance and Physiological effects to see if it is possible to identify physiological signals of performance decrements and associate them to the variation of a specific factor. The validation of the concept would provide a solid support for the redesign of HMI / procedures / training, and eventually for the use of adaptive automation in the cockpit.

Finally, a deep dive analysis of pilots' performance during the simulation allowed the identification of the contextual conditions and factors that led or contributed to degraded performance, and is here used to develop suggestions for HMI improvements and other measures to support the performance recovery.

## 1.5. Structure of the document

This document is composed by three main sections.

**Section 2** is dedicated to the validation of the assumed effect of Scenario 1 runs design on the three factors under investigation (Workload, Stress and Situation Awareness), the assumed effect of runs design on the physiological factors, and the understanding of the impact of the runs on the Pilot Flying performance, giving an initial indication about the actual effect of the combination of the three factors on pilot's performance. Also, an additional sub-section (Section 2.4) is dedicated to the investigation of Pilot Flying situation awareness in the different phases of Scenario 1 through the analysis of eye-tracking data.

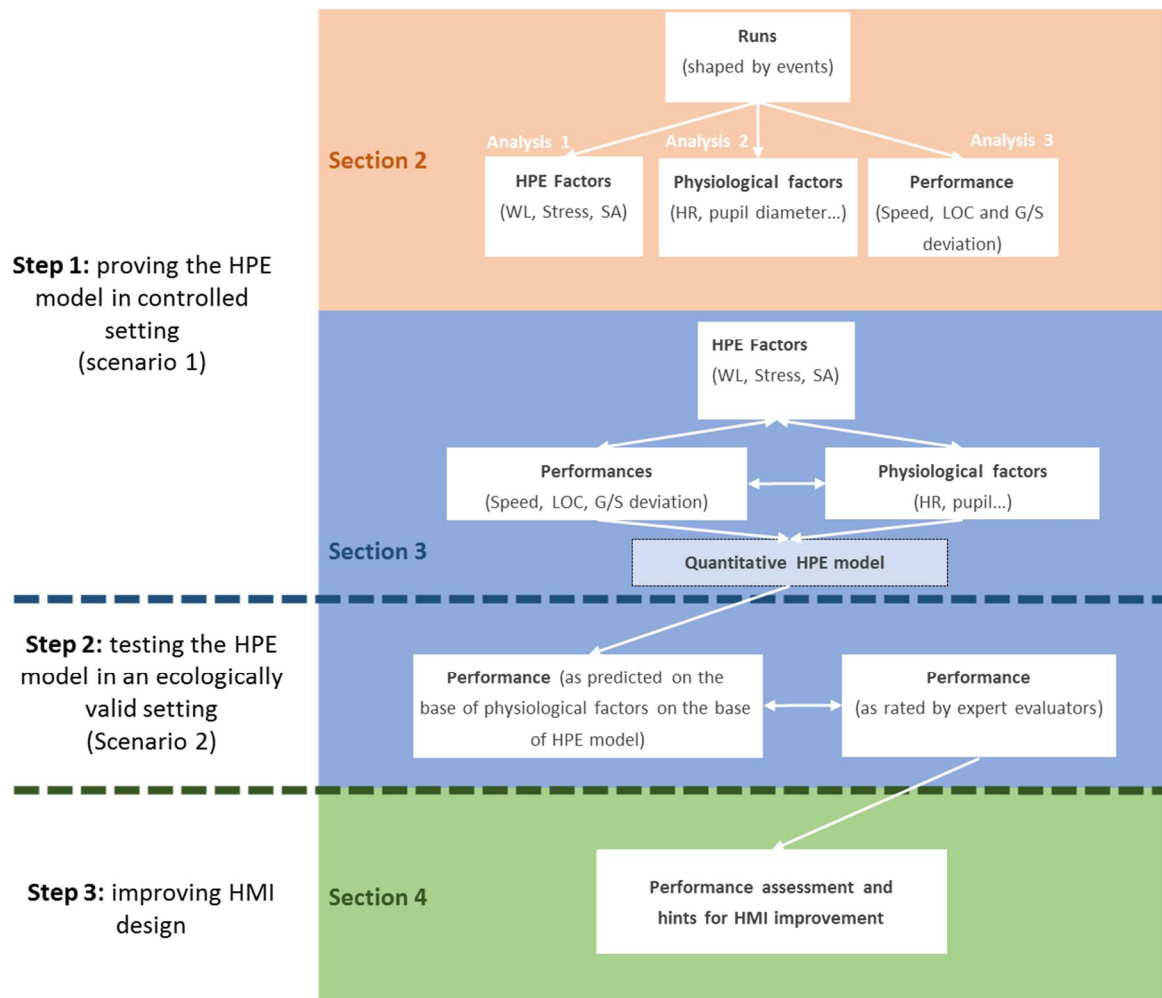
**Section 3** presents the results of a set of correlations performed to validate the HPE concept. The first three correlations – 1) correlation between HPE factors & Performance, 2) correlation between HPE factors and Physiological factors, 3) correlation between Performance and Physiological factors - aim at validating the HPE concept itself using data from Scenario 1. The fourth correlation task applies the results of correlations 1, 2 and 3 to Scenario 2, deriving a predicted performance that is compared to the actual pilot's performance assessed through the competence evaluation tool.

Finally, **Section 4** illustrates the ideas for HMI improvements to support performance recovery, based on the analyses of debriefings and cognitive walkthroughs. These results will be used to develop the new HMI to be used in the final round of simulations that will be held in Thales at the end of 2017.

The different sets of analyses performed in each deliverable section are represented in Figure 1.

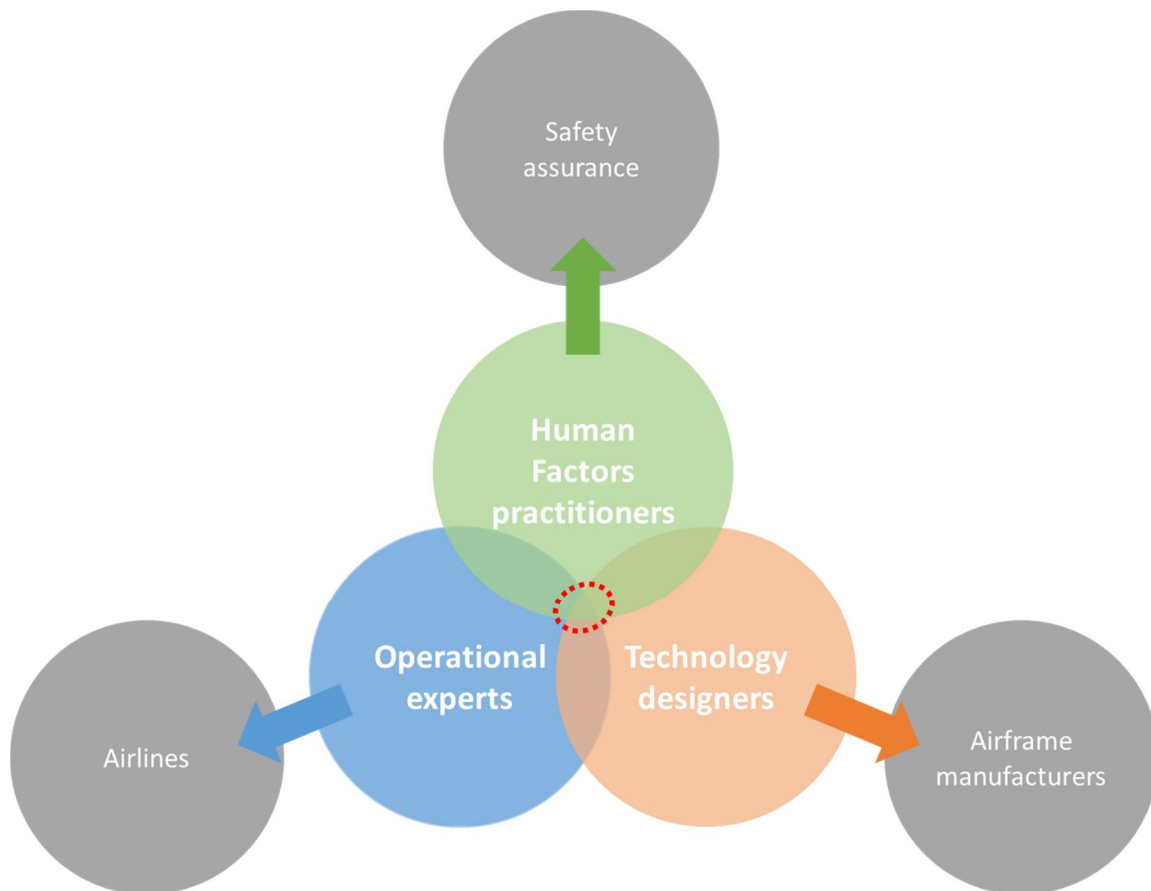
Complex research simulations involving realistic scenarios, full-scope simulators and line pilots, are fairly rare, and attract interest from different researchers asking different questions. Not surprisingly, the first experiment was therefore a multi-stakeholder affair, with a number of intersecting parties coming together to achieve a common goal, albeit with their specific objectives. Figure 2 illustrates these

different research perspectives, with the red circle representing the area of common interest resulting from their intersections. The different actors involved in the collection and analysis of the data were in charge of playing an intermediation role, to achieve a balance in addressing the needs of both project and external stakeholders. This means that the document reflects the combination of this variety of needs, in which long-term research goals are pursued in parallel to shorter-term operational and industrial goals.



**Figure 1: Structure of the document**





**Figure 2: Overview of the research perspectives adopted in this document for data collection and analysis and external stakeholders**

## 2 FACTORS EFFECT ON THE ENVELOPE

### 2.1. Global effects of Workload, Stress and Situation Awareness on the envelope

This section presents the analyses performed on the effect of the different runs of Scenario 1 on three different set of variables associated to the HPE model. The runs of Scenario 1 were performed in the context of the real time simulations held at DLR premises in Braunschweig in May 2016. More details on the simulations can be found in D6.3 “Test report preliminary testing with system pilots’ cognitive task analysis”.

Considering that all the presented analyses will continuously refer to the different runs, a description of the run characteristics is a necessary information. A table summarising the main features of the runs is provided here below. All the runs are based on flying a manual approach to a German airport. The experimental subject was always the Pilot Flying (PF), while the Pilot Monitoring (PM) was a confederate pilot.

**Table 1: Scenario 1 runs, HPE factors supposedly affected and airport**

Run n.	HPE factor intended to be affected, at what level	Events used to affect the HPE factor	Airport to which the approach is flown
Run 1	None		Frankfurt (EDDF) RWY 25L
Run 2	Workload – medium	Medium turbulences throughout whole run	Hannover (EDDV) RWY 27R
Run 3	Workload – high	High turbulences throughout whole run	Frankfurt (EDDF) RWY 25L
Run 4	Workload – very high	<ul style="list-style-type: none"><li>- High turbulences throughout whole run</li><li>- Approach and RWY change during initial approach (between IAF and FAF)</li></ul>	Hannover (EDDV) RWY 27R
Run 5	Stress – high	<ul style="list-style-type: none"><li>- Low fuel situation throughout whole run</li><li>- Delay vectors during initial approach (between IAF and FAF)</li><li>- Loud noise during final approach (between FAF and landing)</li></ul>	Frankfurt (EDDF) RWY 25L

Run 6	Situation awareness – highly reduced	<ul style="list-style-type: none"> <li>- Low visibility throughout whole run</li> <li>- Localiser interference during final approach (between FAF and landing)</li> <li>- Wind shift during final approach (between FAF and landing)</li> </ul>	Frankfurt (EDDF) RWY 25L
Run 7	Workload, Stress and Situation Awareness – all medium	<ul style="list-style-type: none"> <li>- Medium turbulences throughout whole run</li> <li>- Low fuel situation throughout whole run,</li> <li>- Delay vectors during initial approach (between IAF and FAF)</li> <li>- Low visibility throughout whole run</li> <li>- Localiser interference during final approach (between FAF and landing)</li> </ul>	Hannover (EDDV) RWY 27R
Run 8	Workload, Stress and Situation Awareness – all high	<ul style="list-style-type: none"> <li>- High turbulences throughout whole run</li> <li>- Low fuel situation throughout whole run,</li> <li>- Delay vectors during initial approach (between IAF and FAF)</li> <li>- Loud noise during final approach (between FAF and landing)</li> <li>- Low visibility throughout whole run</li> <li>- Localiser interference during final approach (between FAF and landing)</li> <li>- Wind shift during final approach (between FAF and landing)</li> </ul>	Frankfurt (EDDF) RWY 25L

The first analysed set of variables consists of the **HPE factors**, measured through subjective ratings (e.g. NASA-TLX for Workload, SACL for Stress and SART for Situation Awareness). Analysing the relation between the various runs and the HPE metrics is needed to understand if:

- The run events assumed to impact a certain HPE factor are actually impacting that (i.e. *“Does a high level of turbulence - i.e. run 3 - increase workload for real?”*) and how;
- The run events assumed to impact only a specific single HPE factor are instead impacting more than one HPE factor (i.e. *“Does a high level of turbulence - i.e. run 3 -increase workload only?”*) and in case how.

Therefore, this first analysis is aimed at validating the assumed effect of runs design on workload, stress and situation awareness.

The second set of variables under analysis consists of **physiological factors**, measured through heart rate-related metrics and pupil diameter metrics. This analysis is needed to understand what are the

physiological factors impacted by the events of each run and how they are modified (e.g. increase, decrease etc.). In other words, this analysis means answering questions like *"What is the effect of a high level of turbulence on heart rate?"*. For what concerns this second set of analyses, a certain behaviour of the physiological factors is expected, based on the literature review and the pre-test documented in D6.3.

Therefore, this second analysis aims at validating the assumed effect of runs design on the physiological factors (i.e. answering questions like *"Does this run – which was designed with the aim of increasing workload at high level – actually increasing the heart rate metrics in a way that I expect to be reflecting a high level workload?"*).

The third set of variables under analysis consists of **pilot flying (PF) performance metrics**, namely speed deviation and localiser and glideslope deviation.

This third analysis aims at understanding whether the scenario events actually impact PF performance and how. For example, expectations are that a run designed to increase workload at *"high"* level (i.e. run 3) will cause a worse performance than a run designed to keep workload at *"low/routine"* level (i.e. run 1), and that a run design to bring workload to a *"very high"* level (i.e. run 4) will cause a worse performance than the one designed to bring workload at *"high level"* (i.e. the aforementioned run 3). Another set of expectations that is central to the HPE envelope model, is that the run in which HPE factors are supposed to be modified at *"medium"* level (i.e. run 7), will affect performance more heavily than runs in which a single HPE factor is modified, even though that factor is brought to a *"high"* level. For example, run 7 is expected to affect PF performance more than run 3 (where workload only is *"high"*).

It must be noted that the conclusions drawn in the context of this last analysis are referred to the correlation between *"runs"* and *"performance"* only. This can give an initial indication about the actual effect of the combination of the three factors on performance, however, there is no consideration of the correlation with both subjective ratings and physiological factors. A full analysis of the correlations i) HPE-performance, ii) HPE-physiological factors and iii) Performance-physiological factors, is done in Section 3 (see Figure 2).

The three different types of analysis are grouped by HPE factors and their combination. This means that the first part of the section will be structured in this way:

- Section 2.1.1 will present the three analyses applied to the runs in which **workload** is expected to be modified (run 3 and 4);
- Section 2.1.2 will present the three analyses applied to the runs in which **stress** is expected to be modified (run 5);
- Section 2.1.3 will present the three analyses applied to the runs in which **situation awareness** is expected to be modified (run 6);
- Finally, section 2.2 will present the three analyses applied to the runs in which **all the three HPE factors** are expected to be modified (runs 7, 8).

In each subsection, run 1 will always be analysed because of its baseline function.

The remaining subsections will be dedicated to

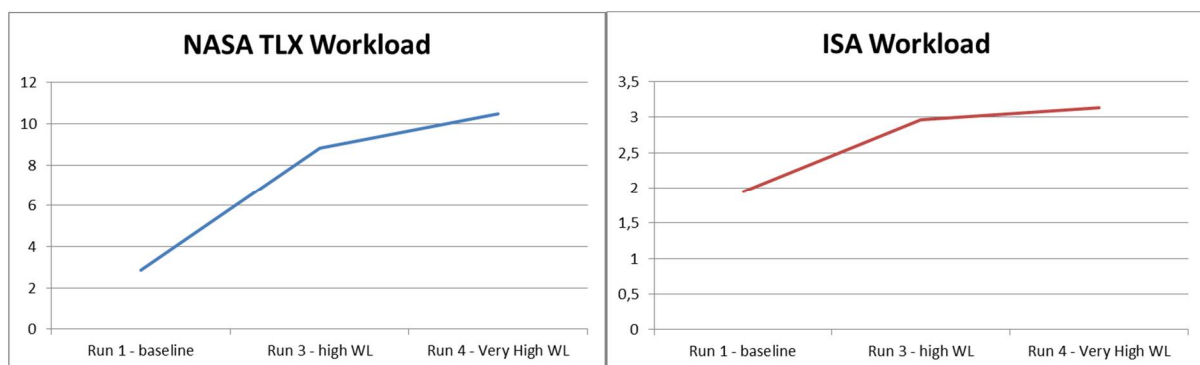
- The analysis of the short-term effect of events impacting stress and situation awareness, namely the loud noise used in runs 5 and 8 to increase stress, and the localiser temporary loss used in runs 6 and 8 to decrease situation awareness (section 2.3)
- The analysis of pilot's Situation Awareness performed through the scan path and eye tracking data (section 2.4)

### 2.1.1. Global effects of workload

#### Workload effects on situation awareness and stress

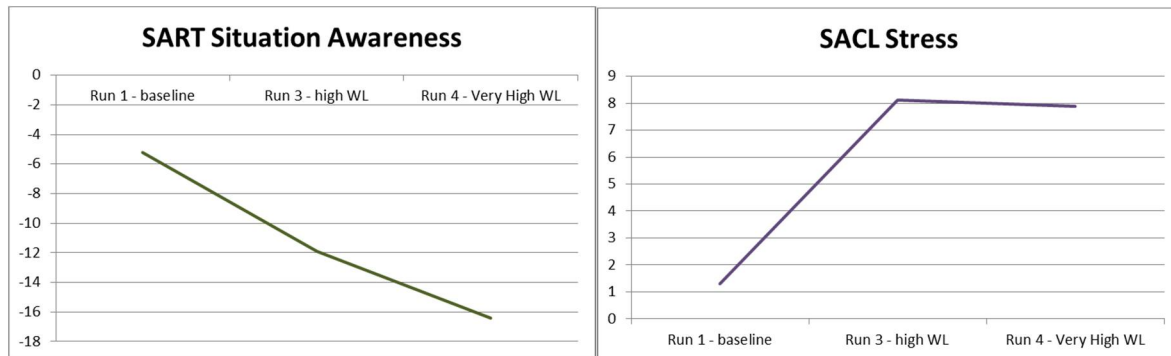
Run 1, 2, 3 and 4 have been designed to increase the crew workload on comparable aeronautical task (landing on an airport). Due to constraints during the experiment, results from run 2 are not usable, but the other runs allow having measures for 3 different levels of task load. This section summarizes these results.

The perceived workload was measured thanks to NASA-TLX questionnaires and repeated ISA measures. Globally, these measures indicate that the workload increased from run 1 to run 4 but that runs 3 and 4 were not different from a statistical point of view. So the approach and runway change does not significantly increase the workload for the whole scenario, even if the increase can be significant for a limited amount of time.



**Figure 3: NASA-TLX and ISA measures of the workload**

It can be noticed that the flight conditions used to increase the crew workload have also impacts on the crew situation awareness (as measured through 10D-Sart) and the crew stress (measured with SACL). Also Workload, Situation awareness and stress were not manipulated independently. The increase of the workload also degrades the crew situation awareness and increase the stress. Nevertheless, the stress level is the same for runs 3 and 4.

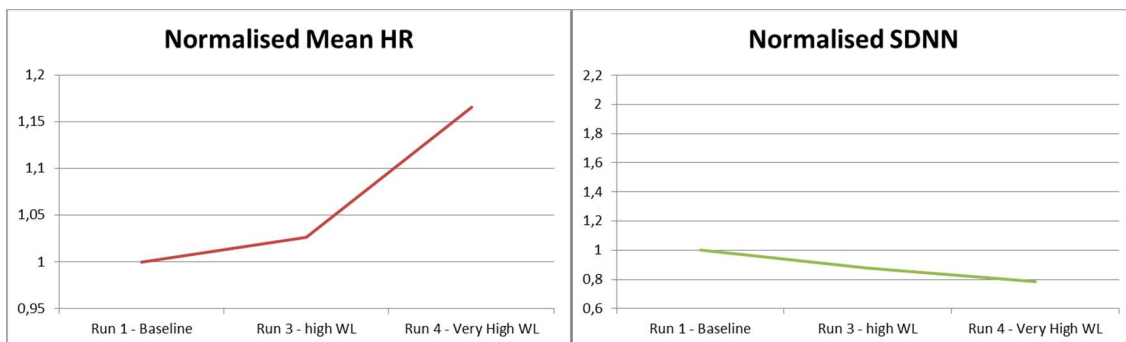


**Figure 4: 10D-SART Situation awareness and SACL stress measures**

In conclusion, runs 3 and 4 can be considered as different from run 1 when workload is considered, but the level of stress is also increased and the situation awareness degraded. Runs 3 and 4 cannot be considered as different from workload and stress points of view, but the situation awareness is more degraded in run 4.

#### Workload effects on physiological factors

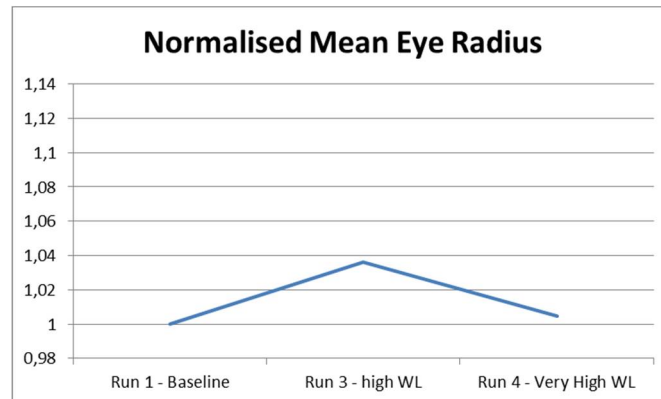
We will focus on normalised heart rate (HR) and standard deviation of NN intervals (SDNN) which are the more relevant measures (see D6.3). Data are kept from the Top of Descent (TOD) to the decision altitude, which allow the comparison of same type of activity for each run (we do not want to include in the analyses go-arounds which are not present in all flights but have a large impact on physiological data).



**Figure 5: HR (left) and SDNN (right) for normalised data in phase 2**

For both measure, the runs are significantly different. Also, the increase of workload (and stress with a decrease of the SA) increases the HR and decreases the HR variability. These results are coherent with the one of pre-tests.

As far as the pupil diameter is concerned, results for normalised pupil radius from TOD to 200ft indicate a significant increase of the radius for run 3 compared to run 1. Result for run 4 has to be taken cautiously, as it relies on very few data, but it confirms a significant increase of the pupil diameter (compared to Run 1).

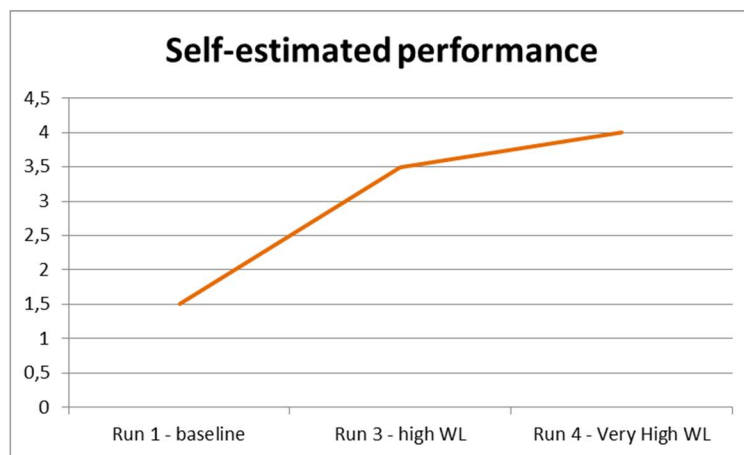


**Figure 6: Normalised pupil radius**

Also, the increase of workload (and stress with a decrease of the SA) increases the normalised pupil diameter (baseline vs high WL or very High WL). Comparison between run 3 and 4 does not reveal an increase of the pupil diameter but we saw previously that the WL level was not significantly different between these two runs. Thus, these results are coherent with the one of pre-tests.

#### Workload effects on self-estimated performances

Performances were estimated by pilots with the use of performance curves (see D6.3). Results show that pilots consider having lower performances when the workload is higher. Nevertheless, once again the difference between runs 3 and 4 is weak.

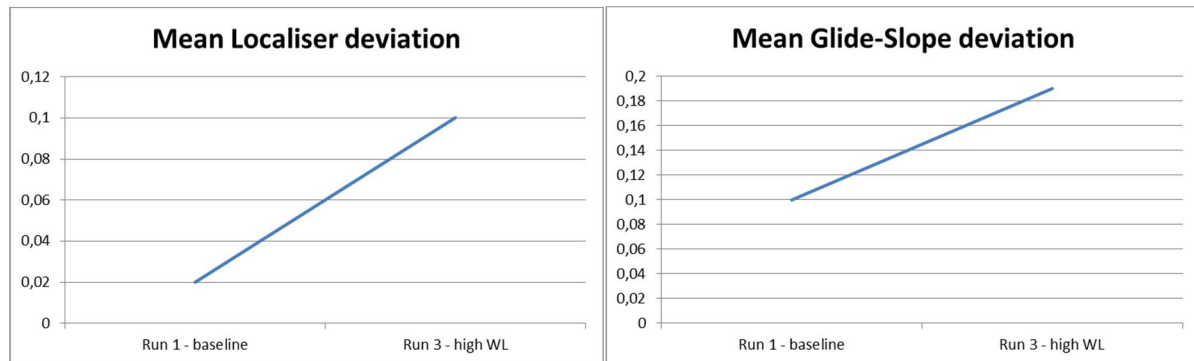


**Figure 7: self-estimated median performance**

#### Workload effects on piloting performances

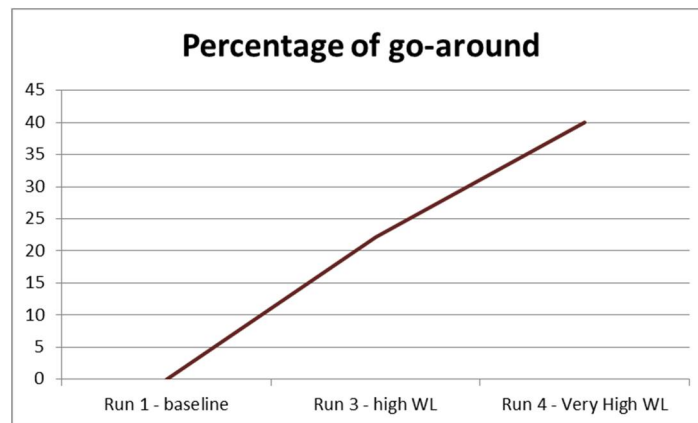
As explained in D6.3, pilot performance was evaluated on the basis of the ability to manually fly the aircraft along a trajectory or along certain target values.

Figure 6 shows that deviations from localiser and glideslope were higher in run 3 than in run 1. Run 4 cannot be compared for these deviations because of the approach change (which prescribe a non-precision approach).



**Figure 8: Localiser and Glideslope deviations.**

It can be noted that no pilot decided to engage a go-around for this baseline run, which strengthen the hypothesis that pilots had acceptable safety margins for run 1 while some of them decided to interrupt the landing and were certainly close to the edge of the envelope for runs 3 and 4.

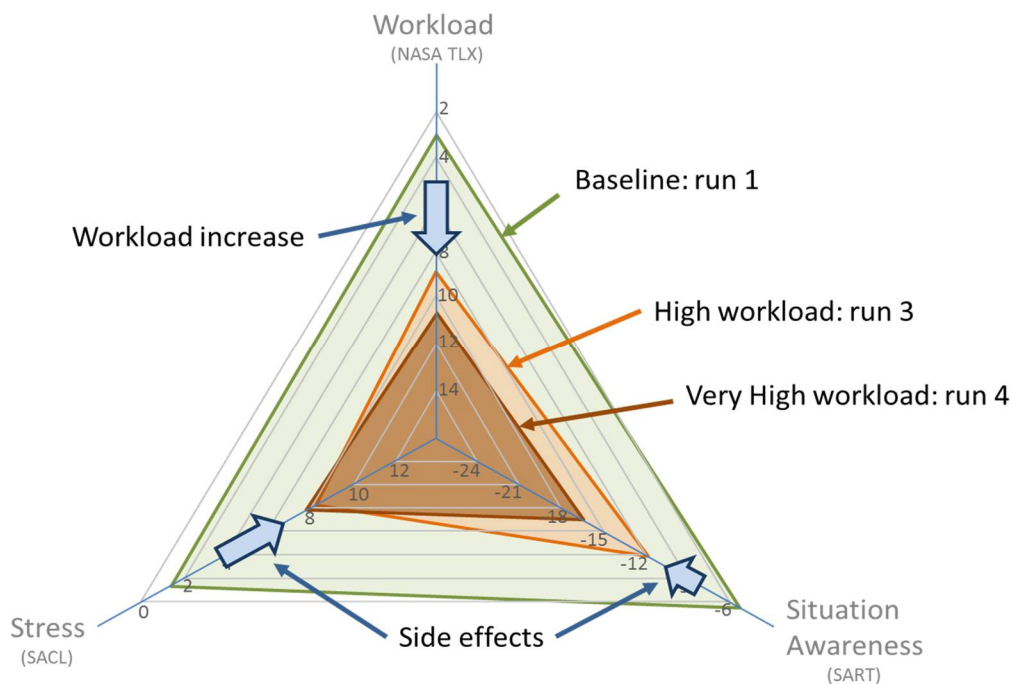


**Figure 9: Percentage of go-around by runs**

### Conclusions about the global effect of workload

The global analysis of runs 1, 3 and 4 demonstrates that ecologic experimentation on flight condition does not allow the gradual increase of workload as it was done in the pre-test. Moreover, the experiment indicates links between three of the components which are supposed to shape the human envelope: the increase of workload is associated with an increase of the stress level and a decrease of the situation awareness.





**Figure 10: Modification of the envelope**

Comparisons with runs 5 and 6 where stress and situation awareness are modified will be used to understand if changes in stress and situation awareness are really 'side effects' and will be more impacted by a direct modification of these factors.

Be what it may, a real difference is identified between the baseline (run 1) and run 3, while runs 3 and 4 have weaker differences (in terms of WL level, stress level, pupil diameter and self-estimated performances). These 3 runs confirmed that the WL increase globally increases the HR, the pupil diameter, and decreases the HRV, the measured performances and the self-estimated performances. Also, the links between physiological markers and the WL, stress and SA levels are confirmed. But these 3 runs do not allow the identification of the edges of the human performance envelope for several reasons:

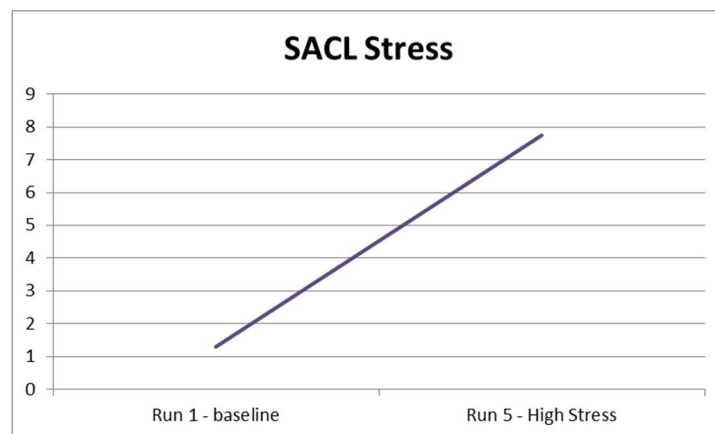
- we do not have an identification of unacceptable performances;
- the task evolves during the run;
- we have only three levels of workload.

Even if this experiment brings no clear evidence of when the envelope was too constrained to let the operator doing the task safely, the study of the number of go-around indicates that the task load increase pushed them closer to the edge of the envelope and some of them decided to interrupt the approach in order to recover higher safety margins. Thus, the results plead for a dynamic adaptation of the envelope encompassing all its dimensions, rather than for independent dimensions with fixed limitations.

## 2.1.2. Global effects of stress

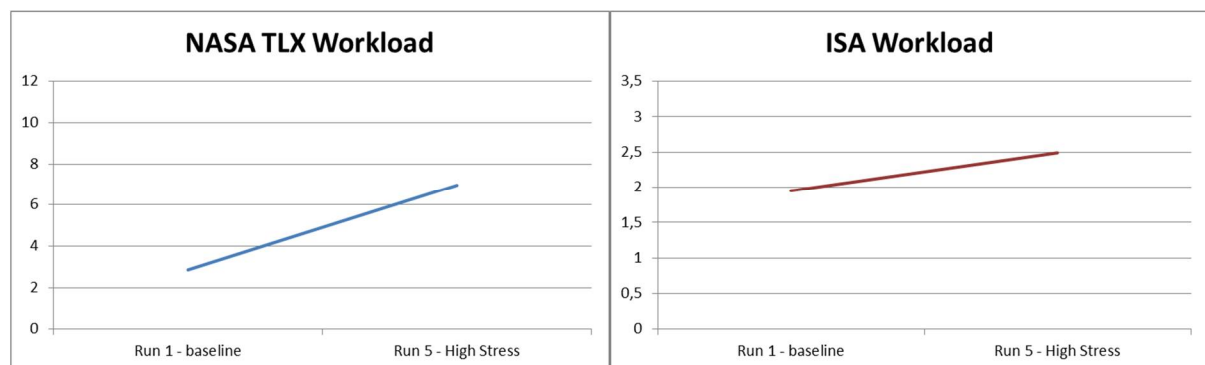
### Stress effects on workload and situation awareness

Run 5 was designed to increase the stress of the crew, with a low fuel situation, delayed vectors and an unexpected loud noise. The SACL Stress measure indicates that these events do increase the stress level.



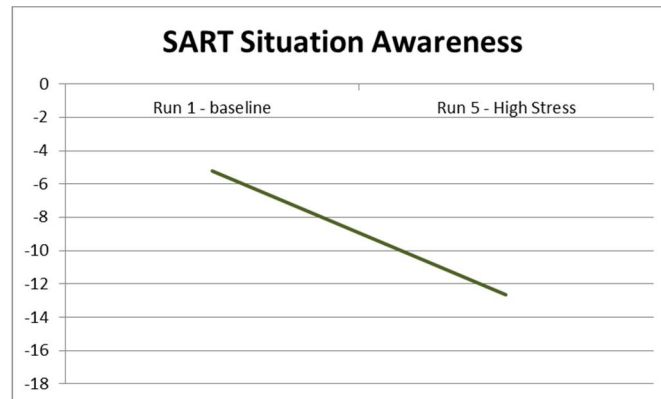
**Figure 11: Evolution of the stress level**

Nevertheless, the stress level obtained is not different from the one given by the workload increase (Runs 3 and 4). Moreover, the addition of these stressors has also an impact on the workload (Figure 12) and the situation awareness (Figure 13).



**Figure 12: Workload**

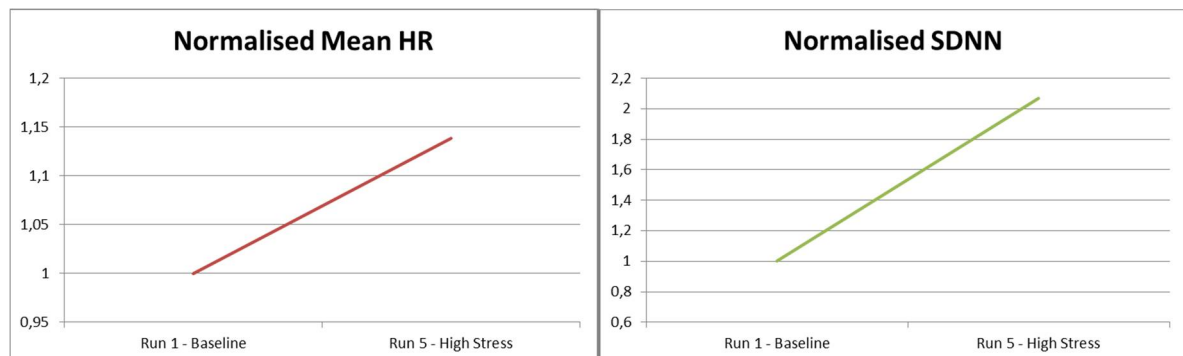
The perceived workload in run 5 is lower than in run 3 and 4, so even in the low fuel situation, the delayed vectors and the loud noise also modify the crew activity and increase its workload, the effect is not exactly the same: we have here the same level of stress but with a reduced workload (compared to runs 3 and 4). Effects on situation awareness are comparable with run 3.



**Figure 13: Situation Awareness**

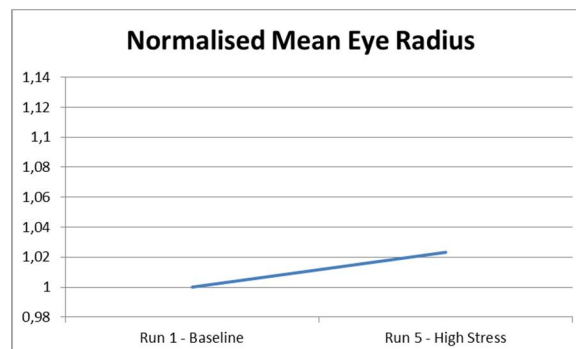
### Stress effects on physiological factors

As expected, the stress increases the heart rate. The standard deviation also increases indicating that the heart rate is more variable in the high stress situation. This result was not expected from the pre-tests, but it could result from a variable level of stress during the experiment: delayed vectors gradually increase the stress level at the beginning of the run while the load noise is around the top of descent and imply a more rapid stress increase.



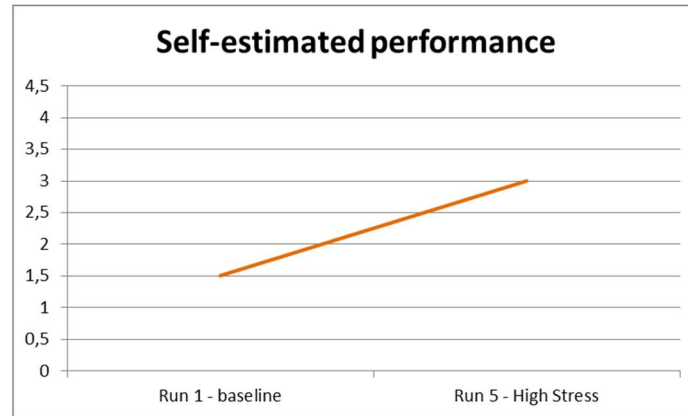
**Figure 14: Normalised HR and SDNN**

As found in the pre-tests, the stress increase is also characterized by an increase of the pupil radius.



**Figure 15: Normalised mean eye radius**

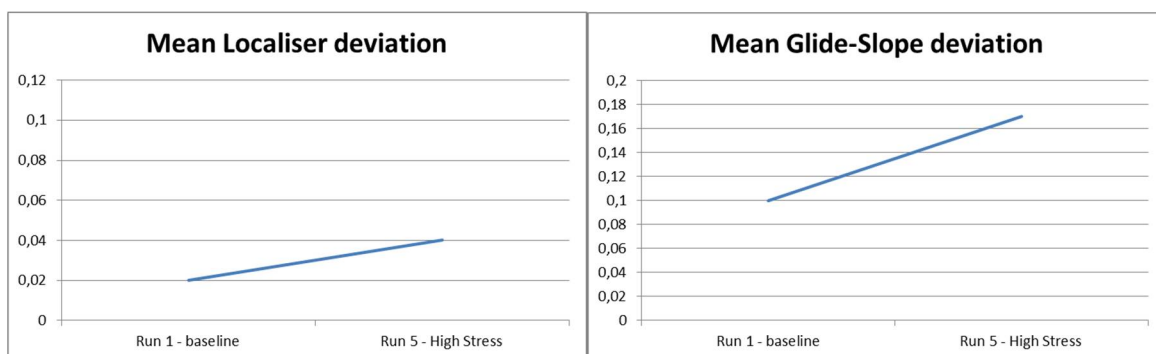
### Stress effects on self-estimated performances



**Figure 16: self-estimated median performance**

Results show that pilots consider having lower performances when the stress is higher, but the degradation is lower than in run 3 and 4. This result is coherent with the indication that the workload level is lower in this run than in runs 3 and 4, with the same level of stress and comparable SA degradation.

### Stress effects on piloting performances



**Figure 17: Localiser and glide-slope deviations**

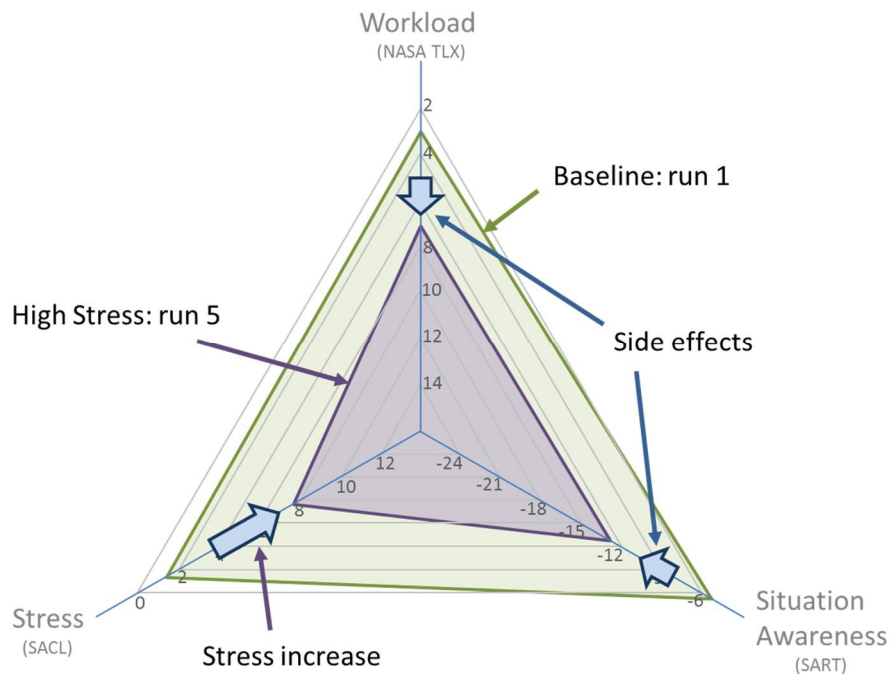
These objective performances show that localiser differences are smaller than in run 3 and 4. This can be explained by the level of turbulences which is lower in this run and directly impacts the performance. Nevertheless, the glide-slope deviation is comparable to the one of runs 3 and 4, indicating a real performance decrease in this run.

The low fuel situation of this runs advocated for not interrupting the approach and so the results cannot be directly compared to runs 3 and 4. No crew decided to interrupt the approach.

### Conclusions about the global effect of stress

Run 5 is clearly different from the baseline. The stress level is close to the one obtained in runs 3 and 4, but the workload level is smaller. Once again the increase of the stress level in the ecological situation reveals to have side effects on workload and situation awareness. Also the effect on the HPE seems to be

more reduced than for runs 3 and 4, which is confirmed by the study of the performances which are slightly better than in runs 3 and 4.



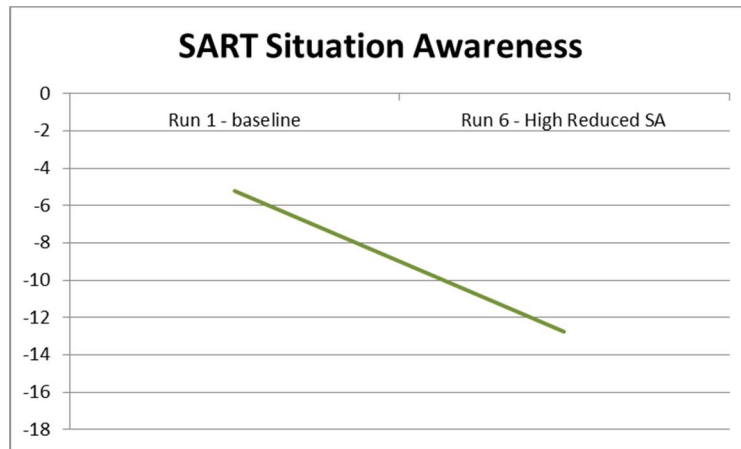
**Figure 18: Modification of the envelope**

This run confirmed that the Stress increase globally increases the HR and the pupil diameter, but decreases the HRV, the measured performances and the self-estimated performances.

### 2.1.3. Global effects of degraded situation awareness

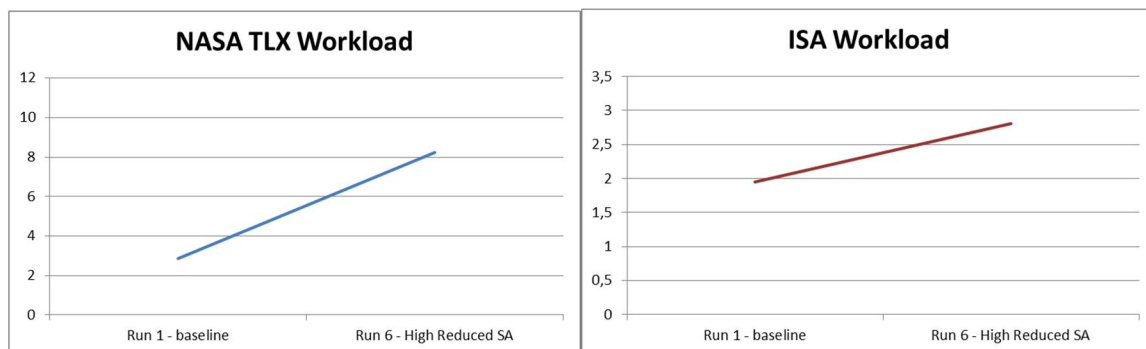
#### Situation Awareness effects on workload and stress

Run 6 was designed to create a high reduced situation awareness, with the combination of low visibility, localiser interferences and a wind shift. The situation awareness, as measured by SART is effectively reduced from the one of the baseline (Figure 19).

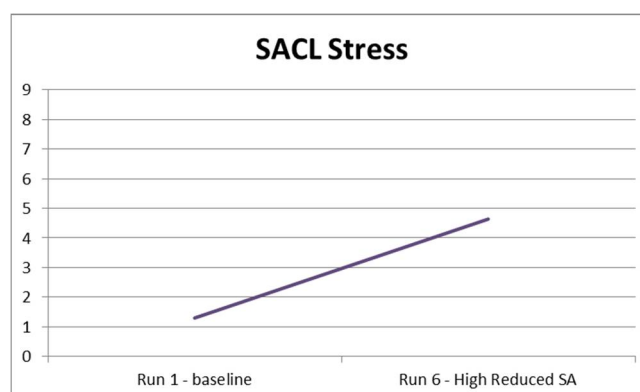


**Figure 19: SART situation awareness**

The level of SA obtained is close from the one obtained in runs 3 or 5 (High Stress) and less reduced than in run 4. Let us now have a look on the influence on workload. NASA-TLX and ISA results indicate that the workload increases, but remain lower than in runs 3 or 4. Nevertheless the WL level is higher than in the high stress condition.



**Figure 20: Evolution of the workload**

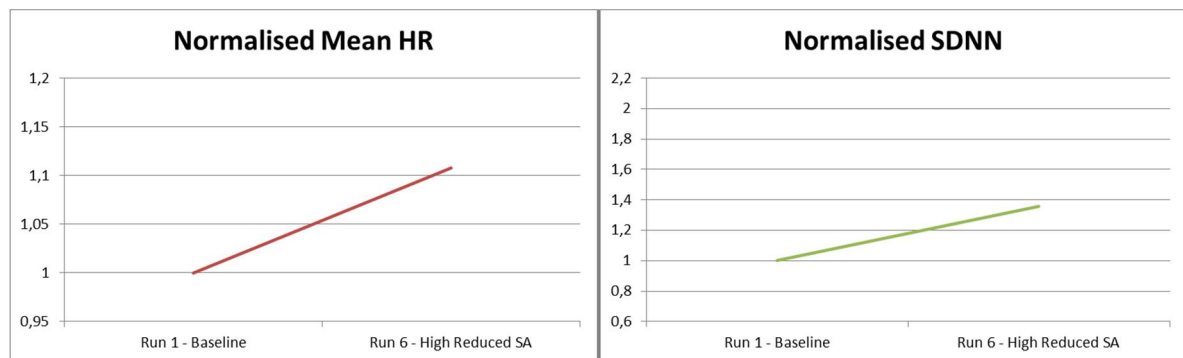


**Figure 21: Stress level**

Figure 21 shows also an increase of the stress level, but smaller than for runs 3, 4 and 5. So the degraded SA comes here with a “small” increase of the stress level and a “relatively important” increase of the workload.

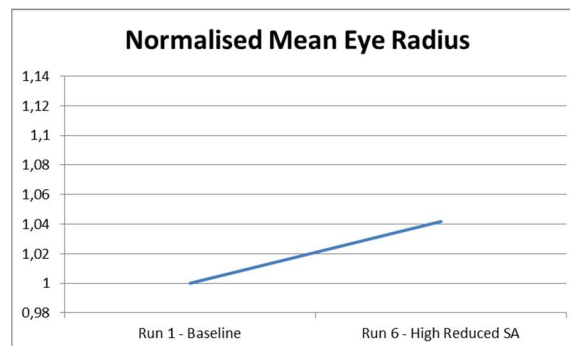
### Situation Awareness effects on physiological factors

Degraded situation awareness comes with an increase of both the heart rate and its variability as shown on Figure 22. The increase is smaller than for the run with high stress.



**Figure 22: Heart rate and heart rate variations**

The normalised mean eye radius also increases in the run with degraded situation awareness, but here the increase is higher than in the high stress’ run.

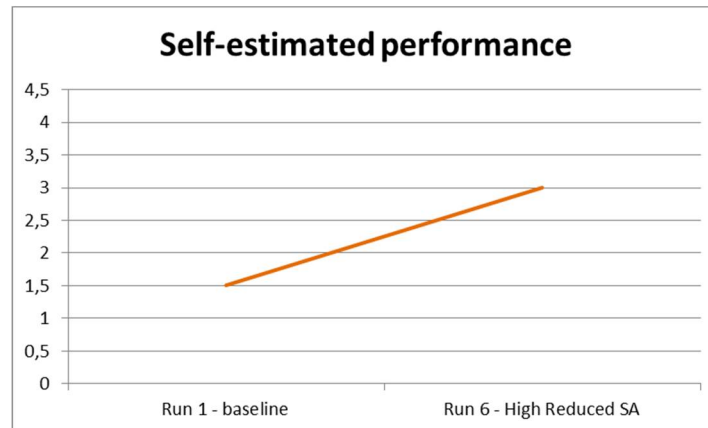


**Figure 23: Eye radius**

The experiment shows an impact of the reduced situation awareness on the studied physiological markers. Nevertheless, as both the stress and the workload levels have also been modified by the experimental condition, the contribution of the degraded SA to these evolutions is difficult to characterise but the results are coherent with the pre-tests.

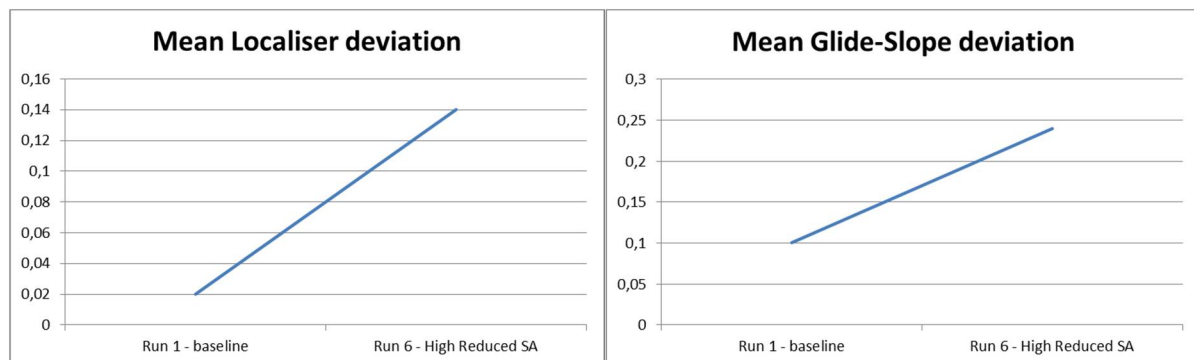
### Situation Awareness effects on self-estimated performances

The degraded SA implies a decrease of the self-estimated performances, at a level comparable with the high stress condition (run 5). So this run is very similar than run 5, with similar WL, SA and self-estimated performances. The main difference is the stress level which is higher in run 5 than 6.



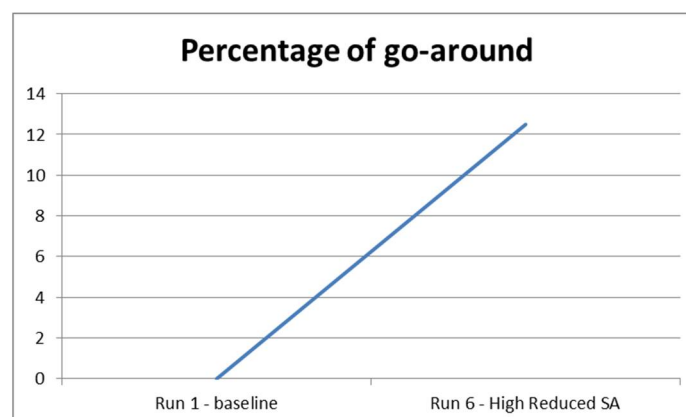
**Figure 24: Self-estimated performances**

#### Situation Awareness effects on piloting performances



**Figure 25: Localiser and glide-slope deviations**

The piloting performances in terms of localiser and glide-slope deviations are lower than in runs 1 to 5. But the values are difficult to compare because the localiser interferences introduced in this run have a direct impact on these deviations.



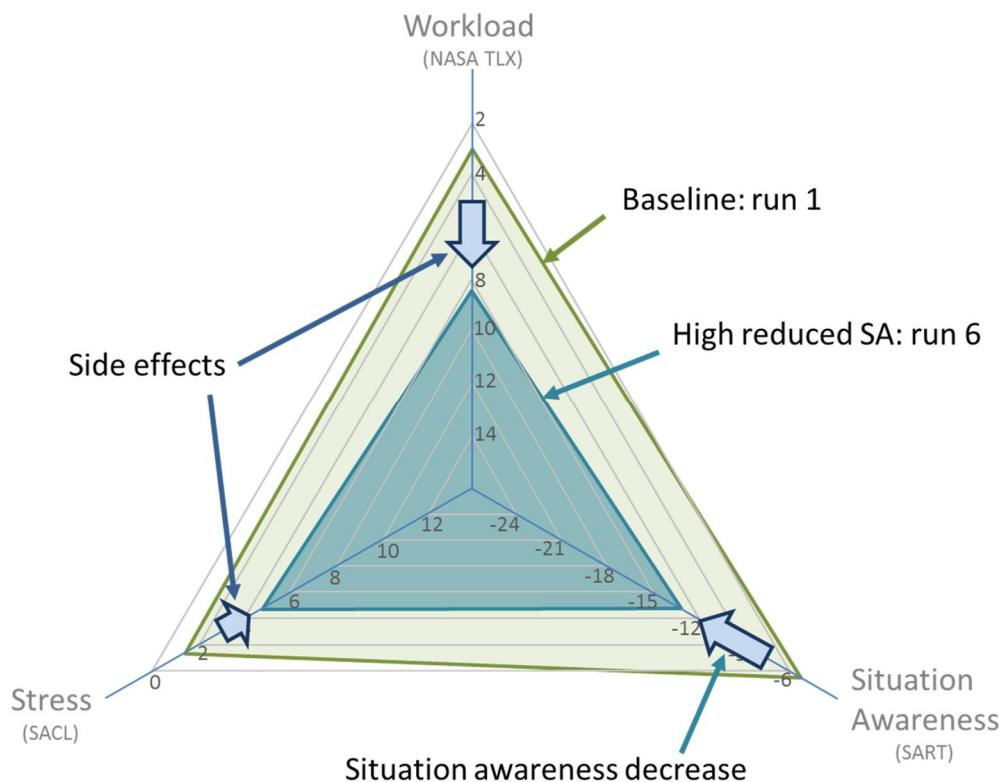
**Figure 26: Go-around**



The percentage of go-around is higher than in run 1 and 5, but smaller than in runs 3 and 4. This result is coherent with results of runs 3 and 4 for which the HPE envelope was more reduced and crews needed more often to recover safety margins by the use of go-around manoeuvres.

### Conclusions about the global effect of degraded situation awareness

The decrease of the situation awareness through a low visibility, localiser interferences and wind shift comes with an increase of the workload and a small increase of the stress level. The resulting envelope is not very different from the high stress envelope, only the stress level is here lower.



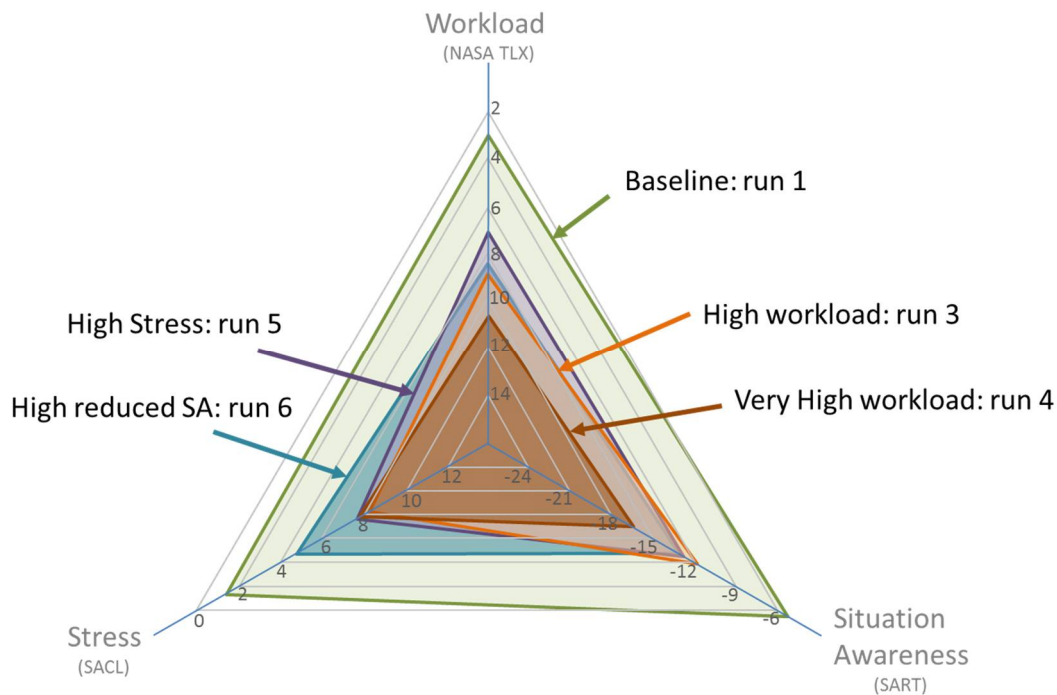
**Figure 27: Modification of the envelope**

The study of performances indicates that these modifications of the envelope induce a decrease of performances and the number of go-around suggests that some crews are at the edge of acceptable performances. The evolution of physiological markers is comparable to the one observed for run 5 (High Stress).

#### 2.1.4. Conclusions about the HP evolution

The experiment was designed to evaluate the impact of the modification of workload, situation awareness and stress on the physiological response of the pilot and the performances. Results first demonstrated that in an ecological situation these three factors cannot be modified independently. The modification of the experimental conditions to change the level of one parameter has always side effects on the two

others. Also correlations between each of these three factors and physiological measures cannot be calculated with these data.



**Figure 28: WL, Stress and SA modification and resulting HP envelope evolutions**

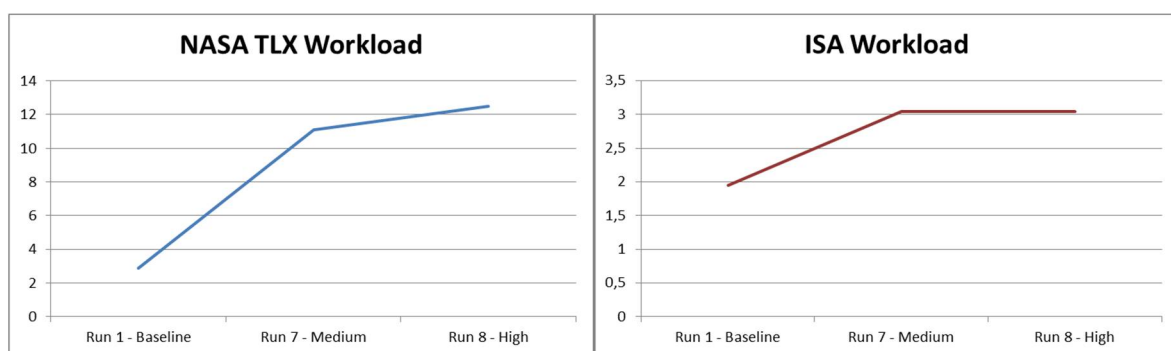
These results on the evolution of the envelope are consistent with the self-estimated performances which are globally worse when the envelope is smaller. The number of go-around also corroborate these results when it is a pertinent measure (run 5 is not comparable because the low fuel situation restrains the possible use of go-around). These results show that reduced HPE pushed some crew at the limit of acceptable performances, and these crews sometimes use go-around manoeuvres to recover.

The reduction of the envelope comes always with an increase of the normalised heart rate and of the normalised mean eye radius. Correlations between each individual factor (stress, workload and situation awareness) and these physiological markers cannot be precisely evaluated with these experiments. It has to be noted that the heart rate variability, measured with SDNN, seems to be modified not in the same way by the workload increase than by stress or degraded SA: while in runs 5 and 6, SDNN increased from the baseline, in runs 3 and 4 where the workload is higher, SDNN has a tendency to decrease compared to the baseline.

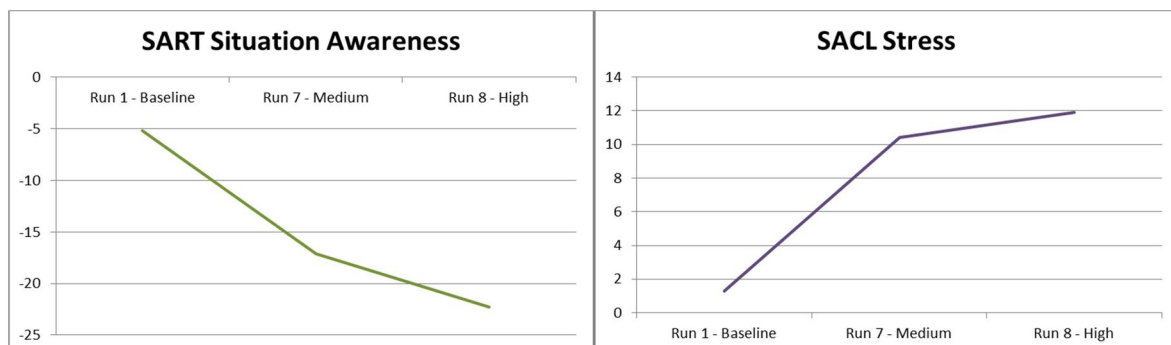
## 2.2. Global effects of combined increase of workload, stress and degraded situation awareness

Let's now study the impact of the combination of factors on the evolution of the HPE and performances. Even if we saw previously that the 3 factors (WL, Stress and SA) were not independent, runs 7 and 8 try to gradually increase the 3 factors to better evaluate the combination effects.

Figure 29 and Figure 30 show that the combined degradation of workload, situation awareness and stress, reduces severely the HP envelope. Even in the medium condition, WL, stress and SA are more affected than in run 4, 5 or 6.

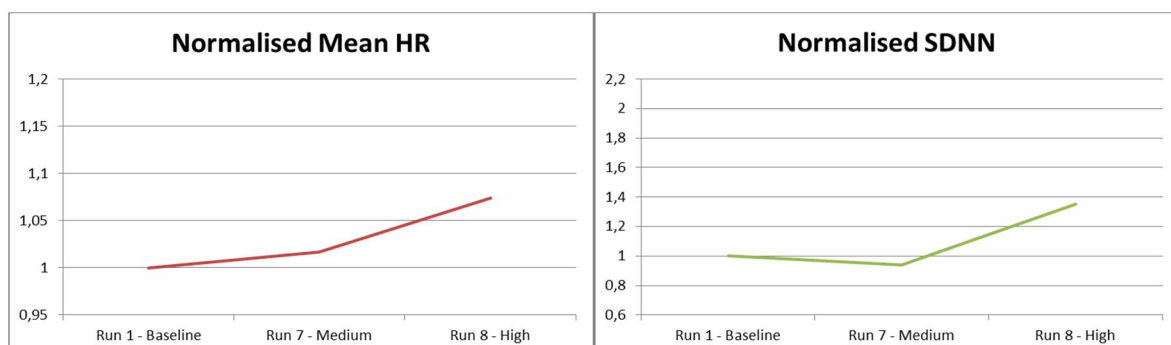


**Figure 29: Workload**

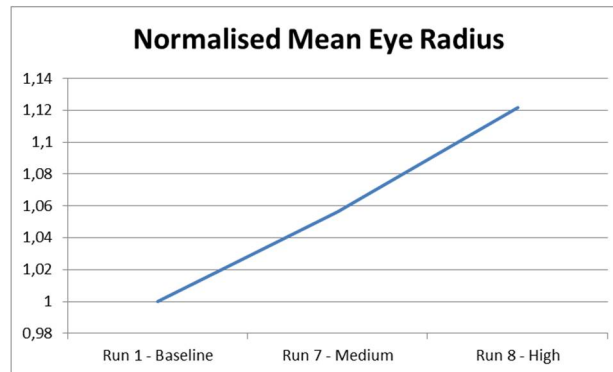


**Figure 30: Situation Awareness and Stress**

### Combined effects on physiological factors



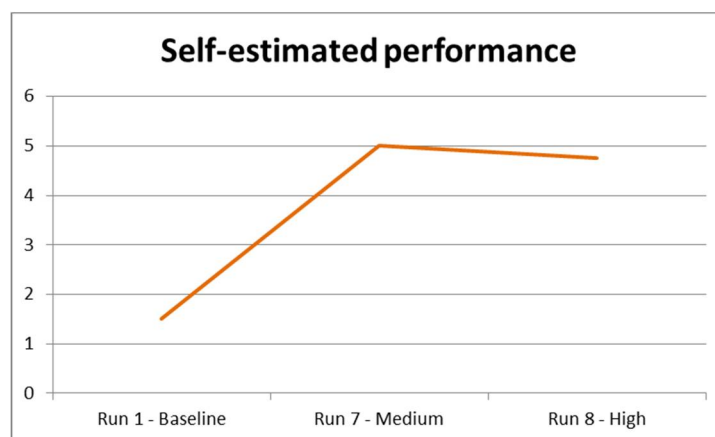
**Figure 31: Heart rate and heart rate variability**



**Figure 32: Normalised mean eye radius**

### Combined effects on self-estimated performances

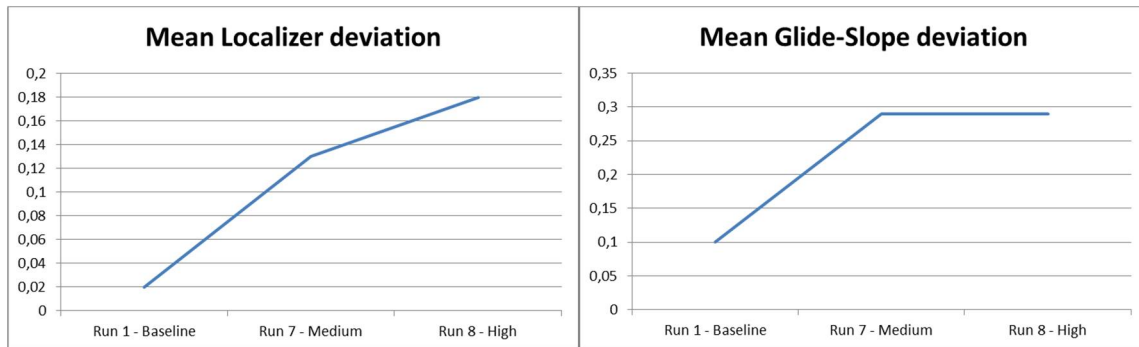
Self-estimated performances indicate a more severe degradation of the performances than in all the other conditions. Nevertheless, there are no differences between the medium and high conditions (run 7 and 8).



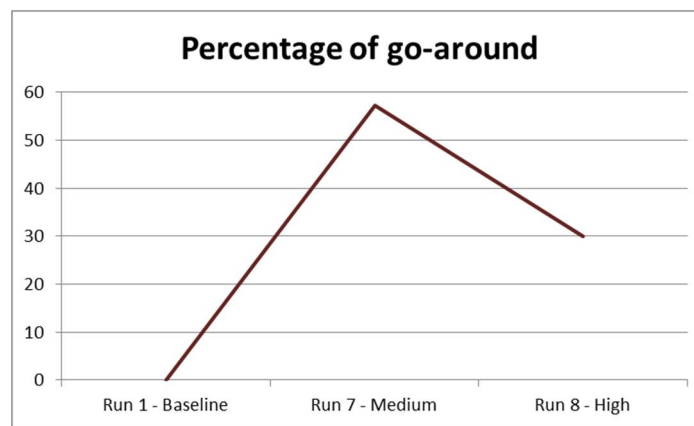
**Figure 33: Self-estimated performances**

### Combined effects on piloting performances

Once again, localiser and glide-slope deviations show worse performances than in the other runs, and the percentage of go-around is the highest of all the runs in run 7. The experimental conditions of run 8, with a low fuel situation were not favourable to a go-around. This fact can explain the reduced number of go-around in run 8 compared to run 7.



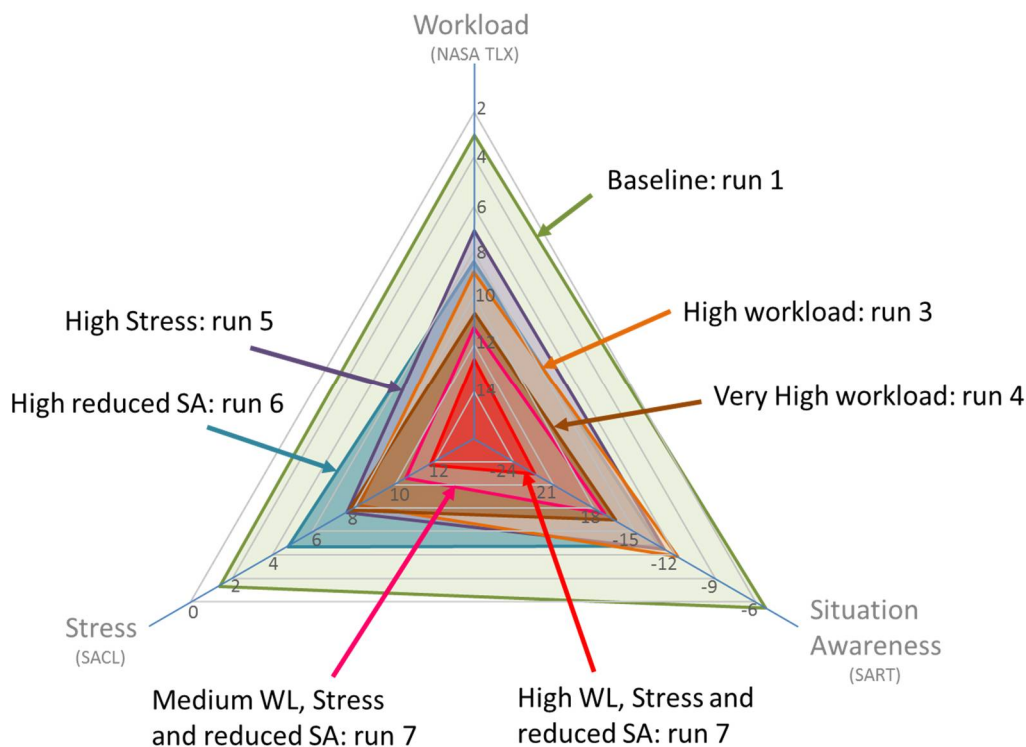
**Figure 34: Localiser and glide-slope deviations**



**Figure 35: Percentage of go-around**

### 2.2.1. Conclusions about the global effect of combined increase of workload, stress and degraded situation awareness

Runs 7 and 8 demonstrated that factors that primarily impact the 3 dimensions of the HPE studied here combined adversely and reduced more severely the envelope. Moreover, the performances seems to be more critically affected, with a very frequent use of go-around manoeuvre to recover safety margins, even if in some case (low fuel situation) it could be a questionable decision.



**Figure 36: Evolution of the HP envelope**

Results on physiological factors do not defend the hypothesis that the combination of WL, Stress and SA factors degrades more severely the situation. In run 8, the normalised mean eye radius increased from 12% (compared to the baseline), which is more than the combination of runs 3, 5 and 6 increases (respectively 3.6%, 2.3% and 4.2%, that is to say a global 10,1% increase). But a contrary result is obtained for the heart rate, with a global 7.3% increase for run 8, compared with 2.6%, 13.8% and 10.7% (total 27.1%).

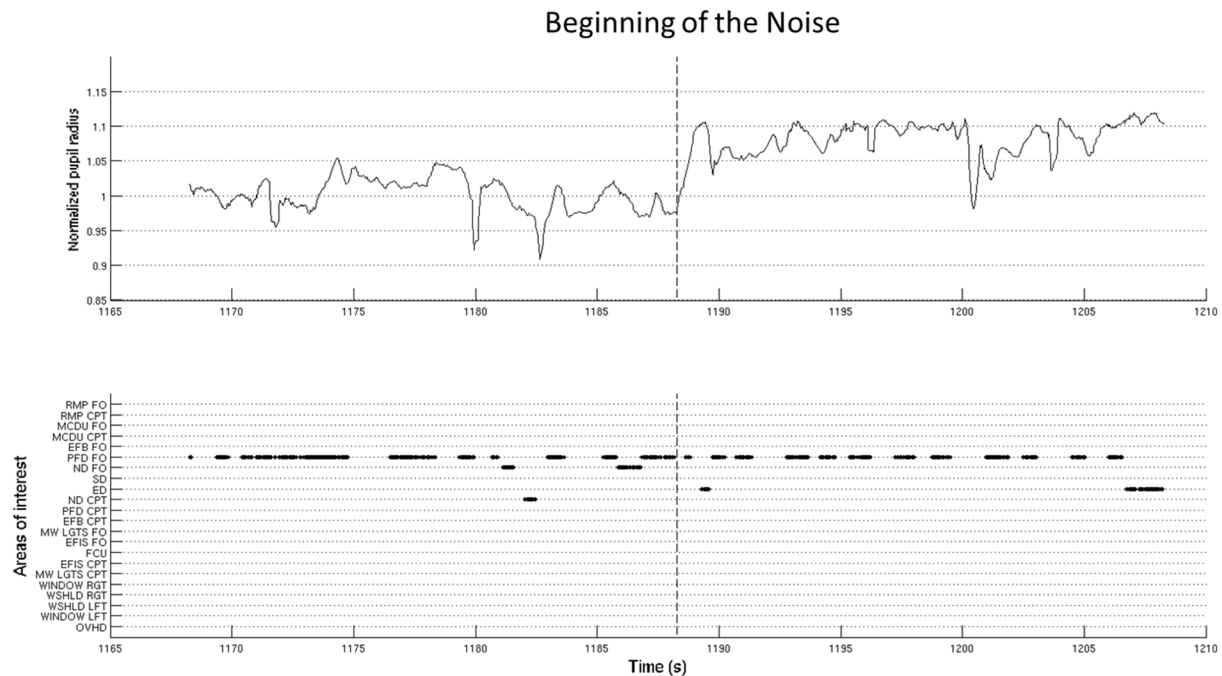
## 2.3. Short term effects of Stress and Situation Awareness on the envelope

This first part of the analysis concentrated on the evolution of the envelope when different levels of stress, workload and SA shape the activity from one run to the other. This allows having a global view of the envelope for each run. Nevertheless, the level of Stress, WL and SA is not constant during each run and we will now try to understand how events that modify these factors influence the physiological markers and the envelope on a short period of time.

### 2.3.1. Short term effect of Stress

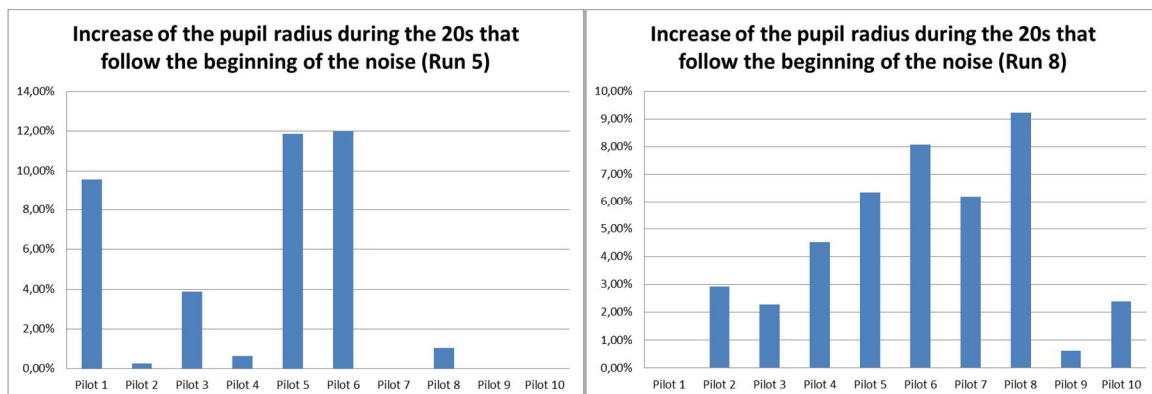
During runs 5 and 8, a loud noise is used to add stress to the crew during the flight. We will now study the impact of this loud noise on physiological factors. The figure below displays the radius of the pupil 20

seconds before the beginning of the loud noise and during the 20 following seconds (upper part of figure 35), as well as the areas of interest watched by the pilot.



**Figure 37: pupil radius and areas of interest around the beginning of the loud noise (Run 8, Pilot 6)**

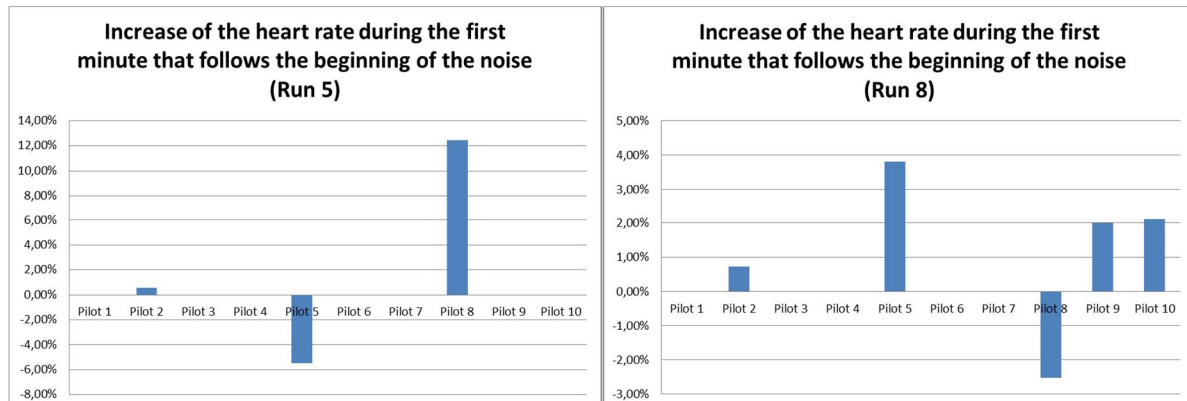
Data of run 5 and 8 indicate that the normalised pupil radius increases during the 20 seconds following the beginning of the noise. The average increase is 5.61% for run 5 and 4.53% for run 8 (see Figure 38 – eye tracking data are not available for all pilots).



**Figure 38: Increase of the pupil radius after the beginning of the loud noise for each pilot and each run where a noise was introduced**

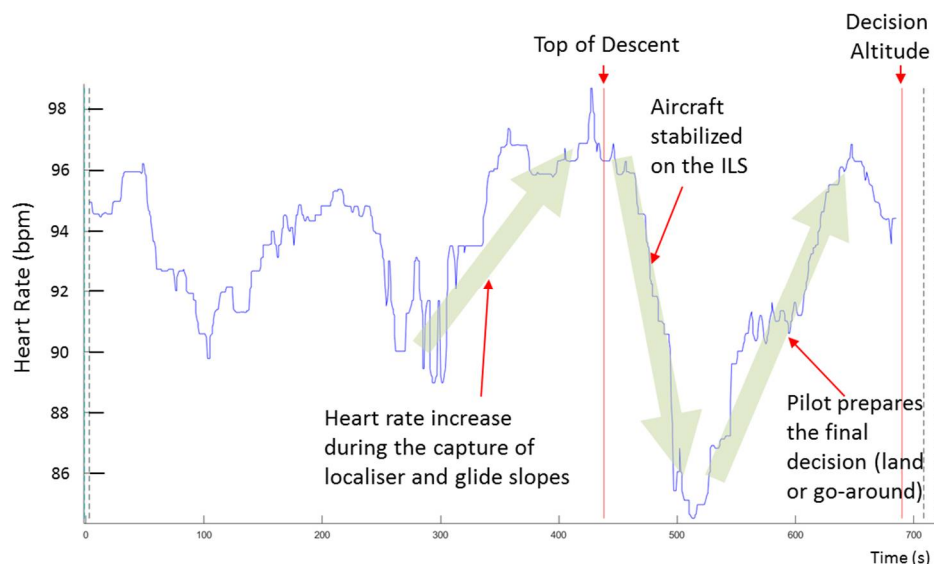
Results for the study of the heart rate variation around the beginning of the loud noise do not show an increase of the heart rate (see Figure 39).





**Figure 39: Increase of the heart rate after the beginning of the loud noise for each pilot and each run where a noise was introduced**

As a matter of fact, modification of the heart rate is a slow process (compared to the modification of the pupil radius) and the short term change implied by the stressor is difficult to extract from the global shape induced by the task. Figure 38 shows the heart rate for a typical flight in baseline condition (Run 1). We can identify an increase of the heart rate when the crew prepares the final approach (just before the top of descent). Then the heart rate decreases when the pilot is following the ILS track and starts to increase again before the crew take the decision to land (here few seconds before the decision altitude). The loud noise arrived during the final approach and only modifies this more important general variation of the heart rate which is linked to the task.

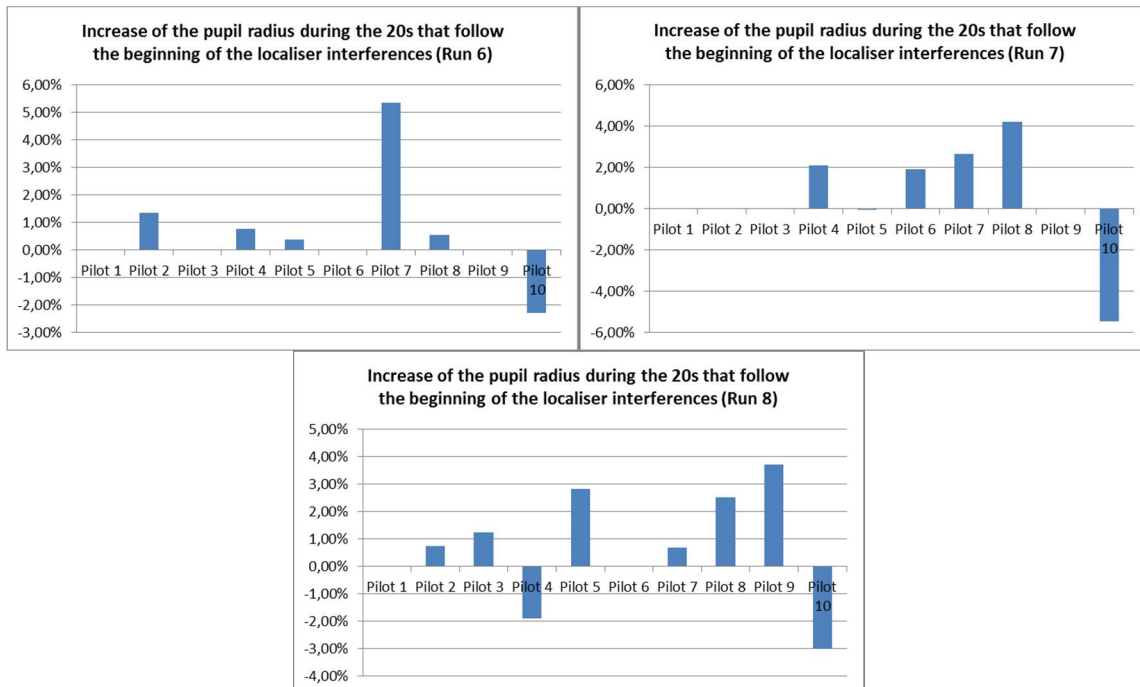


**Figure 40: Typical evolution of the heart rate during the landing flight phase**

### 2.3.2. Short term effect of reduced situation awareness

During runs 6, 7 and 8, localiser interferences are used to reduce the situation awareness of the crew during the final approach. We will now study the impact of this event on physiological factors.





**Figure 41: Increase of the pupil radius after the beginning of localiser interferences for each pilot and each run where localiser interferences were introduced**

Results from run 8 have to be interpreted cautiously because the loud noise and the localiser interferences were not completely disconnected events (the loud noise started 30s to 40s before the localiser interferences and remained during the interferences). The study of the modification of the heart rate just after the localiser interferences is not relevant at this stage of the study as it cannot be easily extracted from the more global pattern shaped by the final approach activity.

### 2.3.3. Conclusions about short term effects

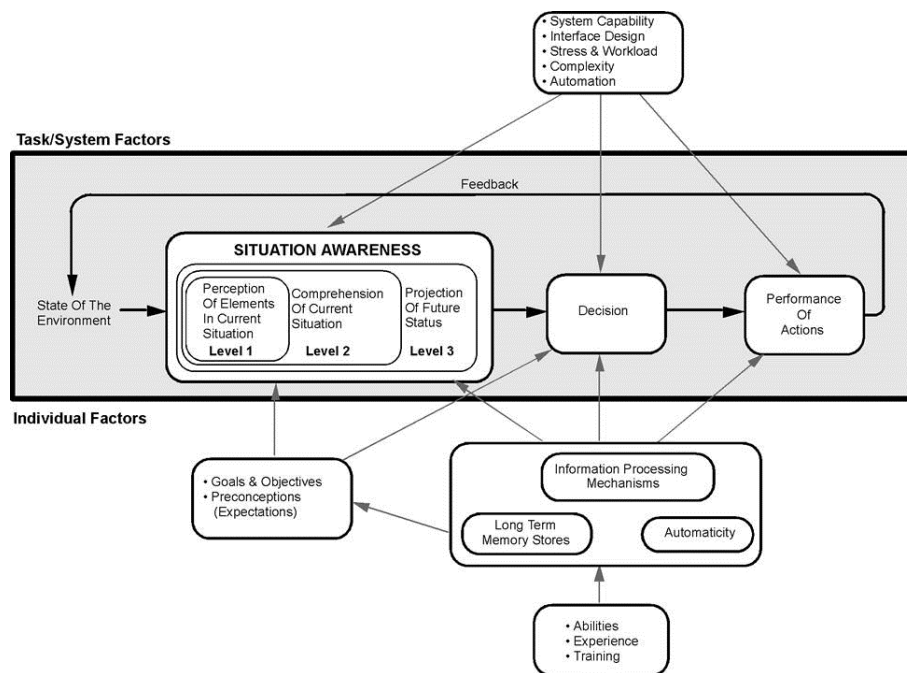
Data from Scenario 1 indicates correlations between one specific event (the beginning of the loud noise) and the evolution of one physiological marker (pupil radius) around this event. Also, the study of sudden changes in the pupil radius could be used to identify some events that shape the HPE. The study of heart rate displays large variations during the experiment but cannot be closely related to HPE factors as we do not have a continuous evaluation of stress and situation awareness levels. Nevertheless, and especially when the situation becomes more and more constrained with time, as in Scenario 2, a study of the evolution of physiological parameters should be more appropriate than a study of the mean value. Results from this study suggest that pupil radius changes could be done with a baseline of less than 30s, while the study of heart rate requires longer durations.

## 2.4. Situation Awareness analysis using scan path and eye tracking data

This section details two treatments of the eye-tracking data acquired in Scenario 1, run 8 of the two-week simulation conducted at DLR in May 2016. The aim of this work is two-fold:

- to understand movement of point of regard in relation to scenario events and
- to begin to understand pilot situation awareness (SA) in response to scenario events through detailed analysis of gaze behaviour. In support of these twin aims, we provide an analysis of pilot point of regard across run 8.

This analysis is informed by the areas of interest (AOI) defined by the validation test plan (see D6.3). Secondly, we present an in-depth, proof-of-concept analysis of the eye tracking data for a single pilot and propose subsequent explanations of pilot SA in response to this analysis. In particular, the emphasis is on performance degradation and points of recovery. To maintain consistency with the project aims and deliverables, the output of this analysis will be framed using Endsley's (1995) three level model of SA (Figure 42).



**Figure 42: Model of situation awareness (Endsley, 1995)**

The eye tracking data for all pilots indicated that they were focussing on the aircraft controls in order to perceive information as indicated by a dwell time of more than 200ms (Yu, Wang, Li, Braithwaite, & Greaves, 2016). However, differences between the pilots were found and these differences may elucidate how information was used to guide decision making. This report makes use of the eye tracking data together with cockpit and eye tracking videos, and SME commentary to start to understand the pilot behaviours providing a proof of concept for future eye tracking analysis used to understand pilot SA.

### 2.4.1. Use of eye tracking for HCI research and design

A central tenet of user-centred design is to understand the user and how they interact with a system. Eye tracking is one method by which the visual behaviour of a user can be understood in greater detail than by observation or interview alone (Jacob & Karn, 2003). Using eye tracking technology, the spatial and temporal characteristics of user visual-behaviour are made subject to analysis and visualisation (Stephane, 2012). These outputs can then be used by a designer to design or modify the tools and systems that a pilot must interact with assuming such systems rely on visual inputs.

Eye tracking methods and their output give a system designer another window through which to understand user response to a system in the visual modality. From the brief summary of key metrics below useful inferences can be made as to where a user is looking, how often they are looking and the overall spatio-temporal scan pattern across multiple complex displays and controls. It is unlikely that there is a set of firm rules or heuristics with which to derive specific interface solutions. However, together with observation and use of multiple methods inference from eye tracking data can assist the designer in identifying areas for development and modification of displays which support an array of tasks. Three basic metrics are described below which can allow certain inferences about user behaviour which may assist the designer.

#### **2.4.1.1. Scan Pattern**

An important element of understanding how a user interacts with visual information delivered by the system is the scan pattern (Ellis, 2009). A scan pattern is the spatial distribution of the acquisition of visual inputs (Glenstrup & Engell-Nielsen, 1995). A user may deploy a specific scan pattern depending on a task. Typically, scan in aircraft cockpits is trained at the private pilot license (PPL) level. At the PPL level the classic 'T' of instruments is introduced and a structured scan is trained to develop pilot situation awareness of the aircraft status (Wickens, Xu, Helleberg, & Marsh, 2001). This classic 'T' has been transitioned into the primary flight display (PFD) in modern glass cockpits and the basic scan is retained albeit within a smaller area. Modern aircraft contain a number of displays distributed throughout the cockpit. For example, the central panel, overhead, PFD navigation display (ND) and the engine and system monitoring displays. Understanding how visual information is acquired between these systems for a given task may lead to insights as to how information should best be located in the cockpit area. For example, if a task demands visual information acquired from dispersed sources necessitating a convoluted scan pattern, this information could be grouped more effectively for that task. A more effective grouping may shorten the scan path and allow for a more efficient synthesis of the information in a single space. Clearly, this may be dependent on the task. However, with appropriate contextual information visual displays can be modified without recourse to changing the physical layout of the cockpit.

#### **2.4.1.2. Fixation duration**

Analysis and visualisation of fixation duration can allow the designer insight into where the more frequently referred to visual information is located (Callan, 2016). As with scan pattern, the most frequently looked at information may indicate the most important or salient information for a given task

(Duchowski, 2007). The system designer can then change the grouping of this information, fuse or otherwise combine this information so that it may be more easily understood or acted upon.

Fixation duration can also give the designer an indication as to the difficulty of a visual task when designing or modifying a system. Fixation durations higher than the overall mean duration may indicate that a user is having to work harder to extract meaning from visual information than might be necessary with a different style of display. Design interventions which fuse the information more effectively or change the way in which the information is displayed may reduce fixation duration and improve performance.

#### **2.4.1.3. Number of fixations**

A high number of individual fixations of shorter duration may indicate inefficient visual search (Salvucci & Goldberg, 2000). If this effect is observed, a designer may wish to change the grouping or salience of information to improve the ability of the user to acquire and search for relevant information in the visual modality (Nakayama & Shimizu, 2004). This may include searching through menu hierarchies or searching for information in electronic flight kits.

This judgment must be made with reference to the task. Conversely, a higher number of longer fixations on a visual display may indicate the relative importance of that display (Jacob & Karn, 2003). The designer may wish to group displays differently or combine information which incurs more frequent fixations

### **2.4.2. Overview of tasks in Run 8**

Run 8 was selected since this run is designed to elicit high workload, high stress, and increased 'low situation awareness.' In addition, run 8 contains the most events compared to the other runs. Run 8 was designed to degrade performance across many factors (workload, or stress, or SA). Run 8 resulted in the highest average workload score of all the runs as measured by the NASA-TLX and the ISA, and the lowest situation awareness score from all the runs in Scenario 1 as measured by SART. The length of the run varied from pilot to pilot, but ranged from 16.1 minutes to 37.2 minutes (mean 23.42, SD 6.74).

In the run, pilots were required to fly an ILS approach with manual control landing at Frankfurt airport, runway 25L. The run starts with increased turbulence which remains throughout the whole run. Three events were introduced to increase stress levels. These events were low fuel, delay vectors and the sudden introduction of a loud noise. The low fuel is an issue from the start of the run. Delay vectors occur from the beginning of the run during initial approach - between the intermediate approach fix (IAF) and the final approach fix (FAF). The loud noise occurs during final approach (between FAF and landing) and lasts for approximately one and a half minutes.

Low visibility is an issue throughout the whole run, localiser interference occurs during final approach (between FAF and landing), and there is a wind shift, from head to tail during the final approach (between FAF and landing). These runs were designed to decrease situation awareness.

### 2.4.3. Performance characteristics of the ideal timeline

This section will detail the elements of the run and the expected actions required by the first officer, who is flying the aircraft. This information was obtained by walking through the run with an A320 SME. This section references figure 6 which shows all AOIs specified by DLR.

The AOI's were specified by DLR at the point of initial analysis. There are 22 AOI's defined within the confines of this project. Anything outside of these areas was deemed to be 'not of interest'.



**Figure 43: Areas of interest**

There is low fuel throughout the run from the beginning, starting with 1780kg of fuel, giving approximately 45 minutes flying time given the weight and type of the aircraft. In order to monitor the fuel level, the first officer (FO) would use the system display (AOI 15) and the engine display (AOI 14). We would expect the FO to focus on either of these panels to obtain fuel information. It would also be expected that the FO would monitor the fuel situation at the start of the run, however, it is also acknowledged that the run begins with the descent preparation, so the background SA that builds during the flight has not been highly developed.

In addition to low fuel, there are increased levels of turbulence from the beginning of the run also. This would require the FO to monitor the primary flight display (PFD) (AOI 17), specifically the speed and the trend information in order to make corrections to the approach. In extreme turbulence, the autopilot can disengage returning full command back to the flight crew. Despite the fact that the autopilot is not engaged during this run due to the FO manually flying, the FO would still need to include the autopilot display in their scan (AOI 8).

Delay vectors occur from the start of the run, for approximately 12 minutes. This would require the FO to monitor the multifunction control display unit (MCDU, AOI 20) and input any required changes.

Low visibility and localiser interference and wind shift would require the FO to monitor their PFD (AOI 17) and the navigation display (ND, AOI 16). The FO may need to compare with the captains ND (AOI 13).

During the loud noise, the desired response would be to check the pressurisation page and the engine parameters using the electronic centralised aircraft monitor (ECAM) (AOI 14 and 15). They would also look at the system display (AOI 15) to assure correct cabin altitude.

The initial analysis used the AOI's (acronym definitions in Table 2) as specified by DLR (Figure 43). To improve the clarity of the analysis, AOIs were grouped according to functions and control responsibility (Figure 44). The five main groups are:

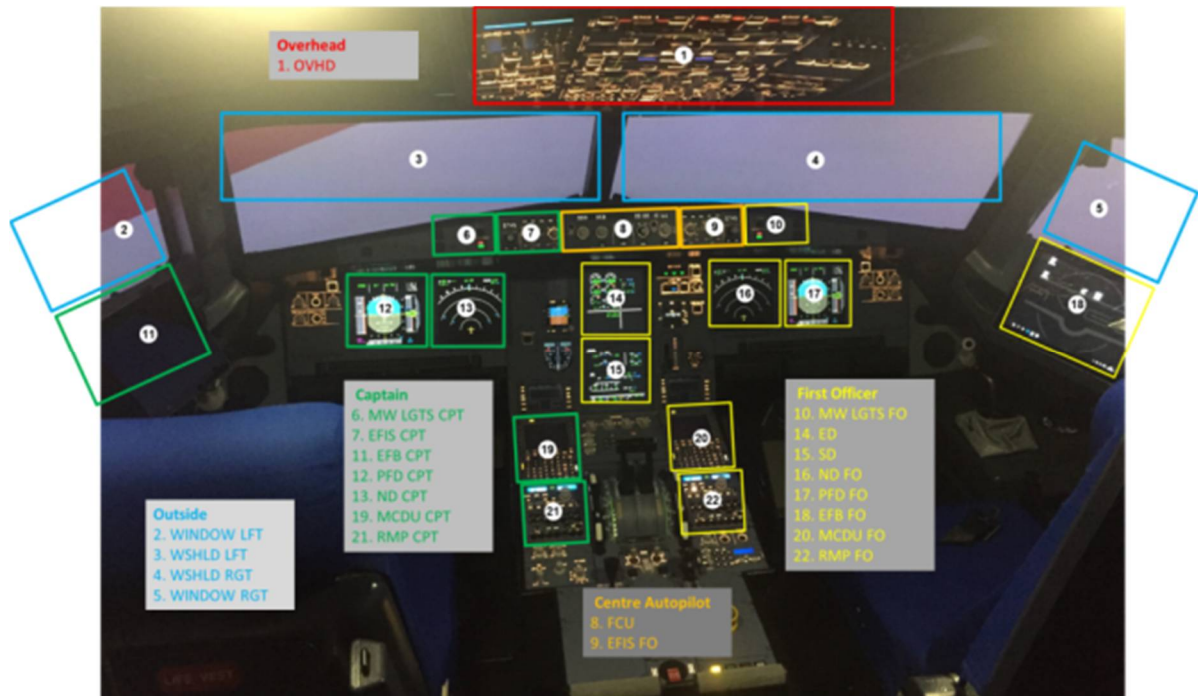
- Overhead - red (AOI 1)
- Outside - blue (AOIs 2, 3, 4 & 5)
- Captain Controls – green (AOIs 6, 7, 11, 12, 13, 19 & 21)
- Centre Autopilot – orange (AOIs 8 & 9)
- First Officer Controls – Yellow (AOIs 10, 14, 15, 16, 17, 18, 20 & 22)

This reduced the number of AOIs and also enabled the analysis to differentiate between the different responsibilities of control.

**Table 2: AOI acronym definitions**

AOI Acronym	Definition
CPT	Captain
ED	Engine Display
EFB	Electronic Flight Bag
EFIS	Electronic Flight Instrument System
FCU	Flight Control Unit
FO	First Officer
MCDU	Multifunction control Display Unit
MW LGTS	Master Warning Lights
ND	Navigation Display
OVHD	Overhead panel
PFD	Primary Flight Display
RMP	Radio Management Panel
SD	System Display
WSHLD LFT / RGT	Windshield left / right





**Figure 44: AOIs grouped by control function**

#### 2.4.4. Method

The details of the simulation can be found in Section 2. This section details the method of the data analysis and interpretation applied to the eye tracking data. The eye tracking data was processed at two levels of granularity: an initial cleanse and analysis to provide timelines for all pilots for run 8 Scenario 1, then a 'proof of concept' deep analysis for pilot 6.

##### 2.4.4.1. Participants

Ten male FOs participated in this study. Seven were German, two Austrian and one French. Participants were between 28 and 36 years old (mean = 31, SD 3.28) and had between 2250 and 7000 hours total flying experience (mean = 4045, SD 1569.23), with at least 250 hours on the A320 (mean = 3125, SD 1557.29) with pilots 3 and 5 gaining over half of their flying hours on the B737. Eye tracking data was unavailable for participant 1. The data included in this report relates to the remaining nine participants (Participants 2 to 10).

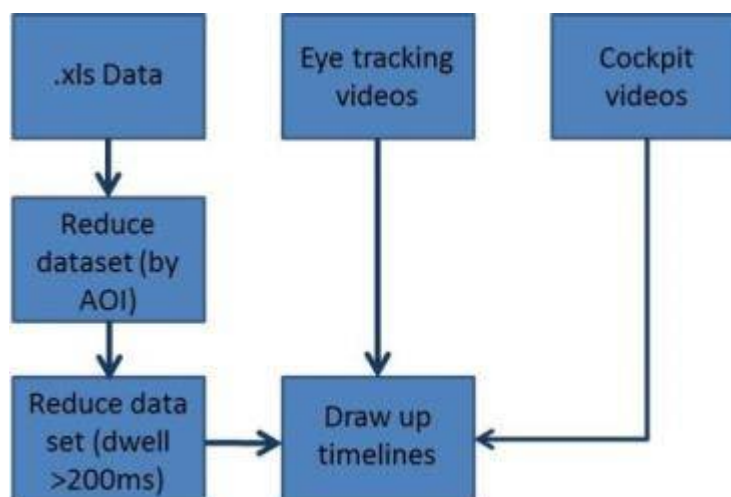
##### 2.4.4.2. Eye Tracking Technology

SMI eye tracking glasses were used to record eye movements during the simulations. SMI eye tracking technology provides binocular tracking at a sampling rate of 120 Hz. Combined with a high definition scene camera and automatic parallax compensation accurate data over all distances can be captured. The SMI BeGaze analysis software supports aggregation of eye tracking data over multiple participants and allows qualitative visualization as well as quantitative analysis of eye tracking data. Data and visualisations such as heat maps or key eye tracking metrics can be exported for further analysis.

#### 2.4.4.3. Approach to Analysis

The DLR Software 'Eye Tracking Analyser' was used for eye-data processing. The tool allows for the analysis for AOIs defined within the simulation environment and for event-related eye data analysis. First, quality metrics for the eye data were calculated. In addition, for each eye data set a validity metric was calculated referring to the percentage of eye data unequal to zero or minus one.

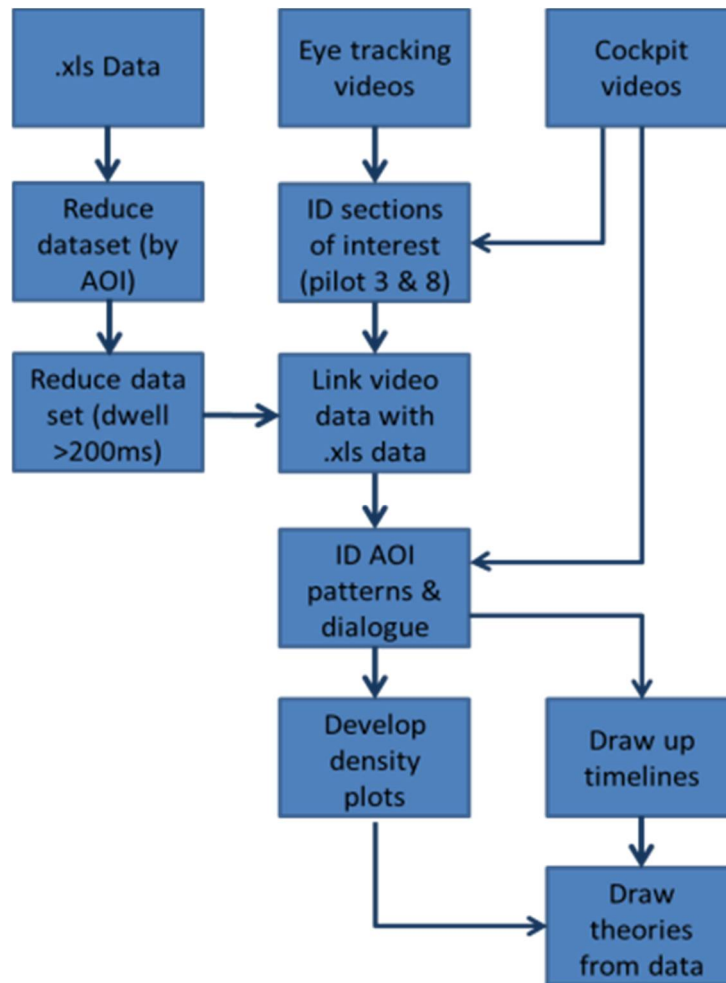
From the raw data fixations, fixations with a minimum duration of 100 milliseconds were calculated. Fixations were then used to compute dwell times. A dwell time is the amount of time the participant paused on an AOI. At this stage the data was given to Cranfield University in Microsoft Excel (Excel) format together with the eye tracking and cockpit videos. Cranfield University then removed the fixations of under 200ms for any given AOI at one time since fixations under 200ms do not suggest that information is being actively processed (Yu et al., 2016). These data were analysed following the process detailed in figure 8 to develop participant timelines described in section 2.4.5.



**Figure 45: Data processing procedure for timeline generation**

For the deep-dive analysis of a single pilot presented in Section 2.4.8, a deeper analysis was carried out for pilot 6 to propose SA insights, following the process detailed in Figure 46.





**Figure 46: Data processing procedure for deep dive analysis**

#### 2.4.5. Pilot timelines

The timelines for pilots 6 and 8 are shown below, in Figure 47 and Figure 48, where each 'dot' indicates a glance. Acronym definitions can be found in Table 2. Figure 47 and Figure 48 show that the pilots spent the majority of their time looking at their PFD. There are, however clear differences between the two pilots, and if they are considered in terms of overall performance, pilot 6 who performed less well, as judged by an expert, only looked at 7 additional AOIs, as well as a larger proportion between AOIs (no defined AOI). Pilot 8 looked at eleven more AOIs, and spent less time looking between AOIs. This may indicate an increased level of SA for pilot 8 when compared to pilot 6, as the FO may have been more aware of the state of the aircraft rather than relying on the captain. In addition, the increased time pilot 6 spent looking between AOIs may indicate less focus on the AOIs and therefore a fluctuation between the defined AOIs. This could potentially lead to less information being absorbed as their focus of attention was more distributed.

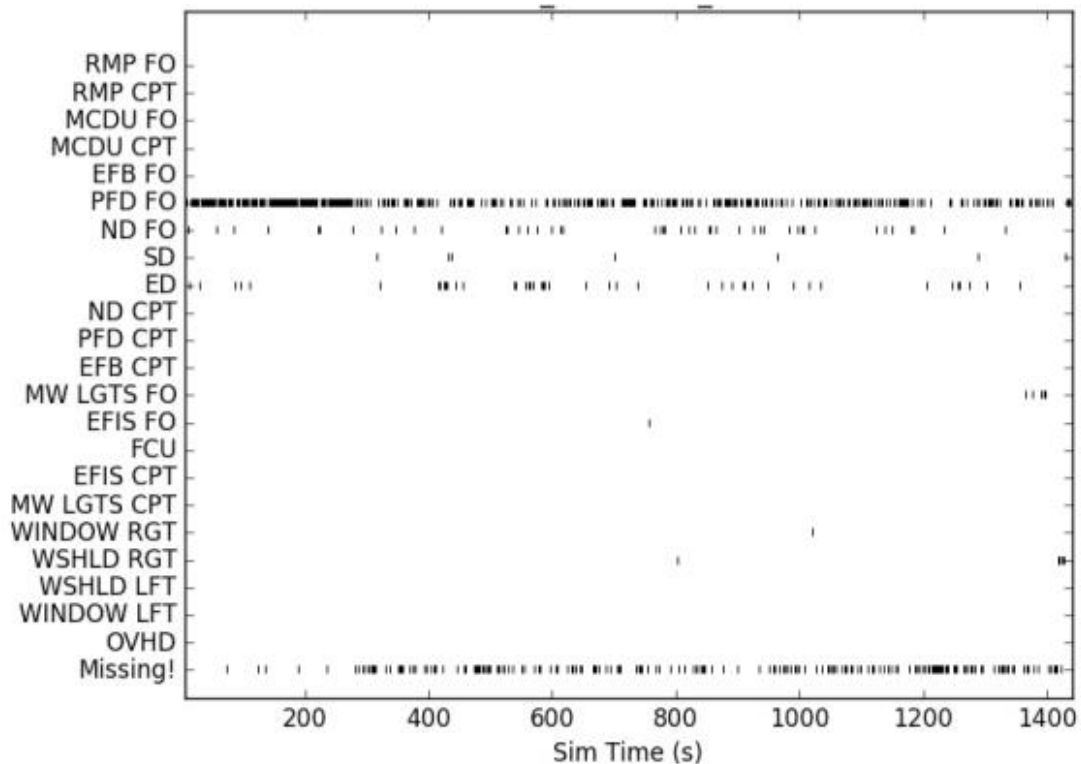


Figure 47: Timeline for pilot 6

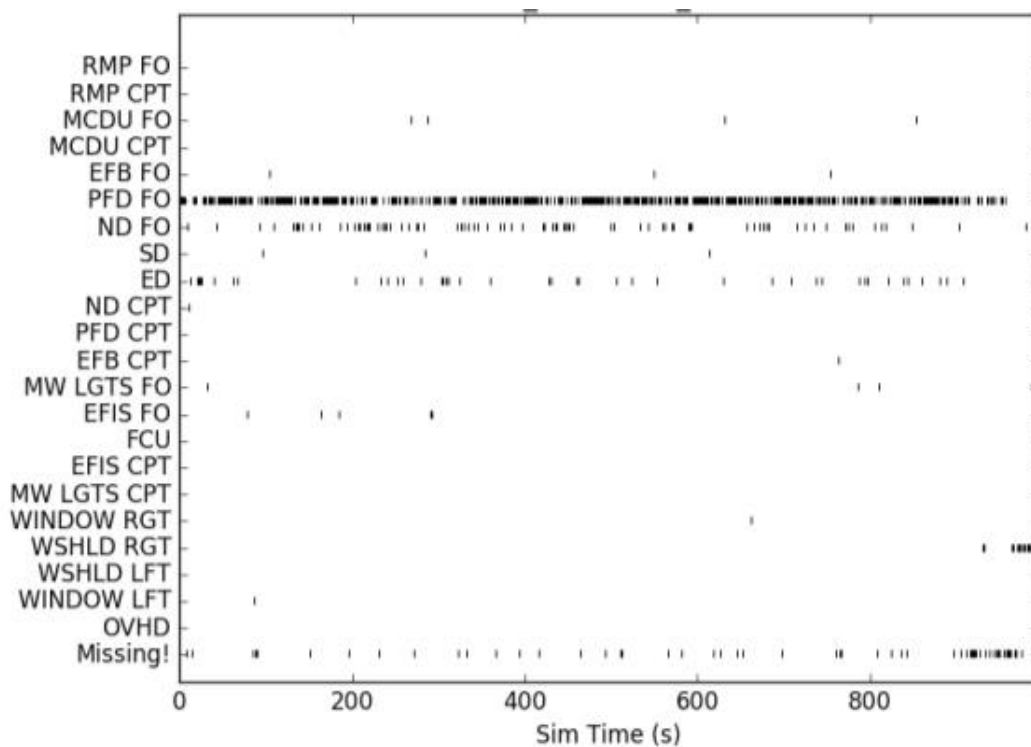
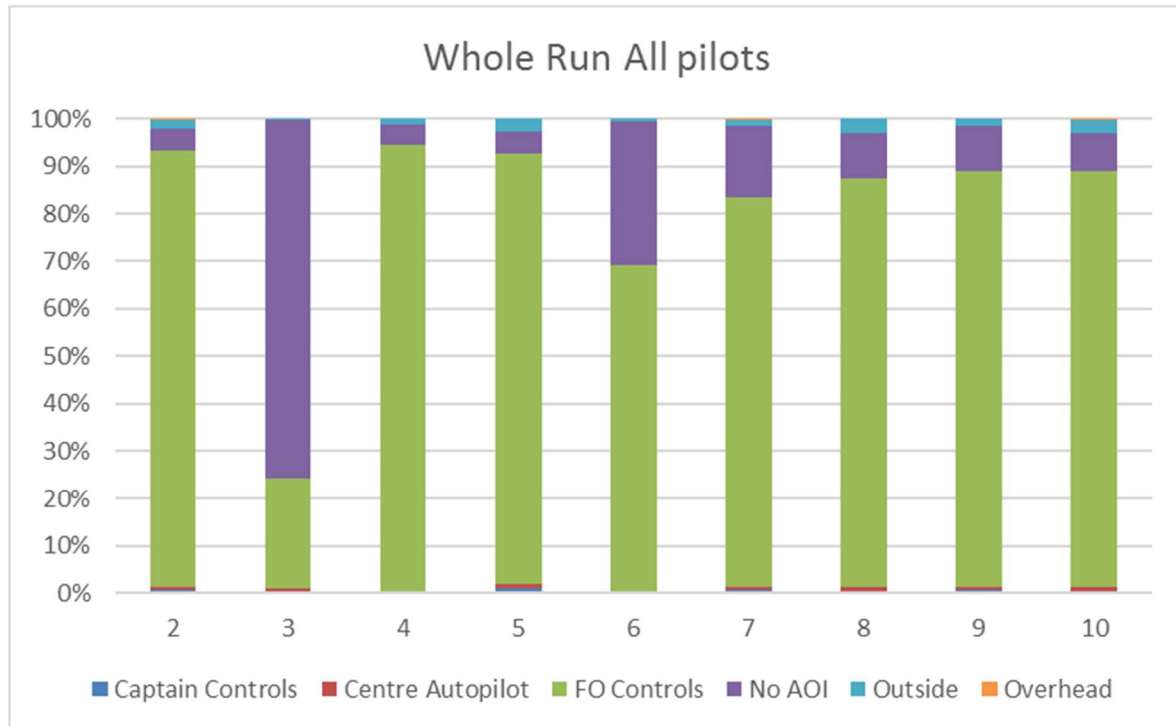


Figure 48: Timeline for pilot 8

#### 2.4.6. Differences between pilots for whole run

A summary plot of percentage dwell time per group of AOIs (as detailed in Figure 44) shows the pilot differences across the whole run (Figure 49).



**Figure 49: Dwell times per AOI for whole of run 8**

The stacked percentage plot in Figure 49 indicates a potential issue with the data quality of pilot 3 which suggests that the pilot spent 75% of their time looking outside the specified AOI's. For pilot 6 this accounted for 30% of their time, whereas the other pilots spent between 4.5 and 15% of their time not focussing on any specified AOI. Pilot 6 was a poorer performer so the higher percentage could be an indicator of reduced SA throughout the run. This explanation is supported by the amount of time pilot 6 spent looking at other areas in and around the cockpit when compared to the other pilots, specifically the captain controls and outside the cockpit. This is particularly apparent when comparing pilot 6 to pilot 8, who was assessed by an expert as a good performer. Pilot 8 may have better level 1 SA since he was attending more to the key AOIs defined.

#### 2.4.7. Differences between pilots for specific events

The following section will look in detail at each scripted event in terms of AOI for each pilot. Stacked percentage charts for each of the scripted events (delay vectors, localiser interference, loud noise and wind shift) can be found in Appendix A.1.

#### **2.4.7.1. Delay Vectors**

During this event, the FO or Captain would be required to input any changes into the MCDU. It was apparent for most of the pilots that the Captain took on this responsibility. In this case the FO may have glanced at the Captain MCDU to make sure that this has been done; not necessarily to check the vectors. This check would be carried out verbally. The plot suggests that the poorer performers, specifically pilot 6 spent less time looking at the captain controls or the centre autopilot. This could indicate that they relied more on the captain to gather input and confirm the vectors; effectively transferring their SA requirements. Clearly there are considerable differences in the scan patterns between pilots.

#### **2.4.7.2. Localiser Interference**

Localiser interference would require the FO to monitor their ND (AOI 16) and PFD (AOI 17) and in addition compare to the captains to ensure that they are aligned. The plot looks much like we may expect, with the majority of time being spent looking at the FO controls. The exception is pilot 8 who in addition spends around 3% of his time monitoring the autopilot.

#### **2.4.7.3. Loud Noise**

During the loud noise, the desired response would be to check the pressurisation page and the engine parameters. Crew would also look at the SD (15) to ensure that the cabin altitude was correct. However, if the FO has been relying on the captain for monitoring purposes, it would be expected that the captain would conduct these checks leaving the FO to fly the aircraft. The stacked percentage plot shows variation in gaze from the FO main controls but not a great deal. Again, this may indicate that level 2 situation awareness is achieved by the captain and the FO is only required to confirm, or ask for confirmation from the captain.

One hypothesis associated with the loud noise scan patterns is that those who acted in a reactionary manner would have scanned the instruments to a greater extent than those who worked with the captain to comprehend the situation to establish current state. There is clearly variation in response and this may be explained by a reactionary versus a calmer approach.

#### **2.4.7.4. Wind shift**

During the wind shift, the FO would be required to monitor PFD (AOI 17) and ND (AOI 16), and also cross-check against the captains ND (AOI 13). During the deep dive analysis it became clear that the wind direction and vector information was handled mainly by the captain. In this instance, the FO would be monitoring their PFD. In addition, the majority of the FOs monitor their PFD and ND during this time, as well as the ED and SD.

### **2.4.8. Pilot deep-dives**

For this analysis a poorer performer was selected. Poor performance may have been caused by a variety of factors. This proof-of-concept analysis may demonstrate how inferences about SA, performance, and recovery measures can be made from the eye-tracking data collected.

The video and Excel data were analysed to provide sections to investigate further by exploring key events and dialogue which could develop understanding of pilot SA during key sections of the run. The flight was also considered holistically to provide an insight into SA across the whole duration of the flight. The aim of these deep dives is to provide a proof-of-concept analytical method for gaining insight into levels of SA from eye tracking data.

#### 2.4.8.1. Key events and dialogue - Pilot 6

Appendix A.2 presents all events and dialogue, taken from the cockpit video for pilot 6. The start time and end time refer to the total elapsed of the run. Time is from run start. The 'heatmaps' (see Figure 50) have been generated using Mathworks MATLAB (MATLAB). The heatmaps display the relative amount of time the FO spent looking at an AOI. The diameter of the red circles on the diagrams are proportional to the amount of time spent looking at the AOI. This time has been normalised to the simulation timeframe. Any time the FO spent not focussing on an AOI has been displayed as arrows if this occurs between AOIs. Any other time not focussed on the AOI's has not been included in the heatmaps. An example of a heatmap can be seen in Figure 50. The remaining heat maps and detailed description of the simulation can be found in Appendix A.3.



**Figure 50: Example of heat-map (AOI frequency and direction pilot 6 from 4:32 to 6:03 minutes)**

#### 2.4.9. Conclusions

The analysis of the eye tracking data and the cockpit dialogue was able to identify how SA was shared between the captain and FO and how this was managed. In most cases, the Captain initiated cross checking with the FO. At a surface level this would indicate that the Captain had better SA than the FO.

However, the FO spent the majority of their time focussed on their PFD, which may indicate a level of shared SA between the captain and the FO, with the FO being supported by their PFD. It was especially apparent in some situations that the FO was effectively 'offloading' their SA to the Captain, with the FO cross referencing information on their instruments when required. Although the eye tracking data cannot explicitly detect performance degradation or recovery strategy, it is able to indicate how the flight crew reacted at key points, for example, when the low fuel situation was realised. This resulted in a significant change in strategy for the flight team, as they then had to manage the low fuel situation. The realisation in the limited fuel level led to the FO and captain working together to establish the future state of the aircraft: Explicit evidence of level 3 SA was captured when the FO was required to project the amount of time remaining given the amount of fuel. The proactive approach of the Captain was different to the reactive approach of the FO. This can be observed through the analysis of the dialogue, supported by the eye tracking data. They managed to recover the situation by sharing information and crosschecking. A certain amount of cognitive processing was also required, in order to calculate the remaining fuel time. Their mis-alignment in views (the captain wanting to call emergency but the FO not agreeing) could have been due to a number of things; the Captain's SA was being supported by external information being fed directly to him, in addition to observing the FO's actions and monitoring the instruments. The FO's SA was supported by the information being fed to him by the captain, along with his own instruments. They were using different information, which as a result built different mental models. It is difficult to envisage how this could be better supported by the interface, but this mis-match indicates that it could potentially be improved; whether this is by interface improvement or SOP changes will need additional analysis.

Evidence of comprehension was reached on more occasions, notably during the loud noise, when the FO was able to establish that there was nothing wrong with the aircraft and that the current situation was normal, by monitoring his instruments. In effect, the FO's SA was being supported by the instruments. For the remainder of the run, the FO effectively offloaded his SA requirement to the captain, who, through communications with ATC and constant monitoring of the instruments may have a more accurate, holistic view of the state of the aircraft than the FO.

This proof of concept has demonstrated that this type of approach to eye tracking analysis can be valuable in giving us an insight into the SA of the eye tracking wearer. This enables us to make certain inferences about the information that is important, and what is comprehended and carried forward.

We believe that this type of analysis does add value, and the combination of the dialogue and the eye tracking data enables some conclusions to be drawn. In terms of carrying this work forward, we feel that a deeper analysis of the existing data would enable firm conclusions to be made. This would include access to the raw eye tracking data to enable us to calculate more accurate fixations and probabilities, and access to the BeGaze software to enable analysis and comparison between the different pilots.

### 3 HPE MODEL VALIDATION

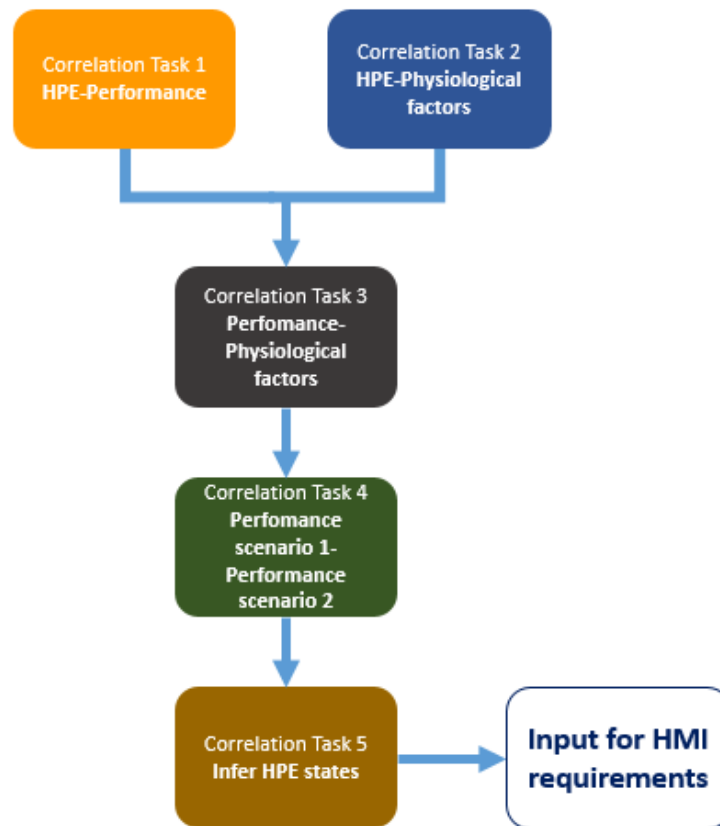
This section is dedicated to the validation of the HPE model. The validation activity consists of a number of correlations that are shown in Figure 51. The validation starts by analysing the correlation between performance and HPE (*Task 1*). Then the connection correlation between HPE and physiological factors is explored (*Task 2*) and finally the two relations are linked in order to analyse the correlation between performance and physiological factors (*Task 3*). All these tasks are performed by using data solely collected in Scenario 1.

In addition to the validation of the HPE model, the correlation activity - if working - can be used to provide a solid support for the redesign of HMI/procedures/training. For example, understanding that a certain performance change is signalled by a specific combination of physiological factors can enable HMI designers to devise triggers for adaptive HMI (e.g. a different visualisation of a specific piece of information). To achieve such an objective, there is a need to verify whether the performance of Scenario 1 can be correlated to Scenario 2 (*Task 4*). In other words, the formula derived as an output of Task 3 is fed with physiological factors collected in Scenario 2 to create a predicted performance. This predicted performance is then compared with the competency performance metrics from Scenario 2 by a regression analysis. On the base of the outcome, it will be possible to infer HPE states (e.g. high workload) from this predicted performance and use this understanding as a basis for HMI redesign.

In summary, the analysis presented in this section will:

- a. Validate the HPE concept (Task 1, 2 and 3);
- b. Connect Scenario 1 and 2 results (Task 4);
- c. Link the HPE concept to the HMI development and evaluation (Task 5).





**Figure 51: Correlation tasks to validate HPE model and predict performance in Scenario 2**

### 3.1. Correlating HPE and performance

The goal of this first correlation task is to determine whether the HPE concept as such actually exists. In other words, the goal is to understand whether the combination of stress, workload and loss of SA lead to a greater decrease in performance than the HPE factors individually.

Therefore, a performance metric will be looked into to see if and how it changes in relation to modifications in HPE factors. The performance metric used for the correlation analysis is the localiser and glideslope deviation. These two deviations are main performance indicators for a manual flight of the final approach (ILS). A single mean value is calculated from these two deviations. This mean value is called flight path deviation and is used for the correlation calculation as the performance metric.

We can define the following variables:

- Variable A the Performance metrics (flight path deviation)
- And Variable B the HPE factors measures, resulting from subjective assessment of Stress, Workload, and Situation Awareness.



The conceptual formula adopted takes in all three factors and adds one interaction-effect variable. The other two-dimensional interaction-factors are omitted because the experimental setup does not provide enough multi-variate sessions to solve a 7-factor equation. Hence it has been reduced to a 4-factor equation. The resulting formula is:

- $\text{Performance}_{\text{FP Deviation}} = C_{\text{ST}} \text{ST} + C_{\text{WL}} \text{WL} + C_{\text{SA}} \text{SA} + C_{\text{Combo}} \text{ST} \cdot \text{WL} \cdot \text{SA} + \epsilon$

Where  $C_{\text{ST}}$ ,  $C_{\text{WL}}$ ,  $C_{\text{SA}}$  are the coefficients of each HPE factor, and  $C_{\text{Combo}}$  is the coefficient for the interactions among them. The coefficients represent the relative contribution of each HPE factor and their combination to the performance. For example, a coefficient of 0.354 for workload means that performance decreases by 0.354 standard deviations if workload increases by 1 standard deviation. The idea is that if the  $C_{\text{COMBO}}$  coefficient is not equal to zero (and the correlation is reliable), then we can prove the HPE concept.

The correlation between HPE factors and performance measures was explored through a multiple regression analysis with interactions. The performance was predicted with the predictors NASA-TLX, SACL and SART and their threefold interactions. The NASA-TLX values were only used as measures for workload as the ISA values did not turn out to be significant in the correlation analysis.

The regression model explains 69% of the variance of the performance ( $F(4,47) = 28.454$ ,  $p < 0.001$ ). The statistical regression equation is:

- **$\text{Performance} = 0.354 \times \text{WL (NASA-TLX)} + 0.285 \times \text{ST (SACL)} + -0.446 \times \text{SA (SART)} + 0.313 \times \text{WL (NASA-TLX)} \times \text{ST (SACL)} \times \text{SA (SART)}$**

Thereby, all predictors predict significant incremental variance of the performance:

- $T_{\text{NASA-TLX}}(1) = 2.643$ ,  $p=0.011$ ;
- $T_{\text{SACL}}(1) = 2.056$ ,  $p=0.045$ ;
- $T_{\text{SART}}(1) = -2.903$ ,  $p=0.006$ ;
- $T_{\text{INTERACTION}}(1) = 2.647$ ,  $p=0.011$

The incremental variance of the INTERACTION term is 4.7%.

**Table 3: Regression table of HPE and performance correlation**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,800 <sup>a</sup>	,640	,618	,04473
2	,829 <sup>b</sup>	,687	,660	,04217

a. Predictors: (Constant), SART, NASA-TLX, SACL

b. Predictors: (Constant), SART, NASA-TLX, SACL, INTERACTION

It can be concluded that the interaction of workload, stress and situation awareness has a significant effect on performance. This proves that the HPE concept is correct and that the interaction of the different factors must be considered and determined when analysing their effect on the performance of pilots. However, it needs to be noted that even though the contribution of the INTERACTION term is significant, it predicts only 4.7% of the variance of the performance. This relatively low value can be explained in part by the fact that the factors are difficult to separate. As reported earlier, it is very difficult or even impossible for example to increase the level of workload without increasing the level of stress. This fact erroneously lowers the value of the INTERACTION term. Therefore, it can be assumed that is higher than calculated in the regression analysis.

Furthermore, it has to be noted that the Performance - expressed as "flight path deviation" - reflects only one part of the performance (from TOD to Decision Altitude), and of course the parameters that express the performance are closely related to the task. So the relationship can only be used for "final approach" with manual handling of the aircraft. Other relevant aspects of the performance, as for example application of procedures and teamwork, are not taken into account in this formula.

### 3.2. Correlating HPE and physiological data

A second correlation task is intended to determine how physiological measures can describe the HPE factors as a proxy-measure to be used in future applications of the HPE model in different scenarios (see Section 3.4).

In order to do that, the HPE factors subjectively measured (Variable A) are correlated with physiological measures collected in real time during the simulation (Variable B). The physiological measures consist of:

- Normalised Heart Rate (HR Norm)
- Heart Rate variability (SDNN)
- Normalised pupil diameter (Eye Norm)

In this case, each HPE factor has its own formula and no interaction effect is foreseen, assuming each factor as an independent measures. The resulting formulas are:

- Stress =  $C_{ST\_A}Phyio_A + C_{ST\_B}Phyio_B + C_{ST\_C}Phyio_C + \dots + C_{ST\_N}Phyio_N + \epsilon$
- Workload =  $C_{WL\_A}Phyio_A + C_{WL\_B}Phyio_B + C_{WL\_C}Phyio_C + \dots + C_{WL\_N}Phyio_N + \epsilon$
- Sit. Awar. =  $C_{SA\_A}Phyio_A + C_{SA\_B}Phyio_B + C_{SA\_C}Phyio_C + \dots + C_{SA\_N}Phyio_N + \epsilon$

We expect that the factors of the formulas will have different coefficients, in other words each factor will be characterised by different configuration of physiological values, as previously discussed in Section 2.

Based on data of Scenario 1, a multiple linear regression approach is used for modelling the relationship between the HPE factors and explanatory physiological variables. For workload, two correlations formulas are calculated, one for the ISA measure of workload and one for the NASA-TLX measure.

Regression output tables are shown for each model created. In all cases the Enter method was used: all predictors were entered into the model. The model ANOVA is reported which, when significant ( $p < 0.05$ ) shows that the model is a better predictor of the dependent variable than the mean of the dependent variable alone. Tables also show the B-co-efficients which are used in the regression equation.

Standardised  $\beta$ -co-efficients are also shown, which can be used to understand the relative contribution of each predictor to the model, independent from the scales on which the variable is measured. We also report the t-tests associated with each predictor. A significant ( $p < 0.05$ ) t-test indicates that the predictor is significantly different to a constant – in other words, that the gradient of the line is significantly different to zero. Adjusted  $R^2$  is reported to control for inflation of R given the number of predictors used in the model. An intercept term is also included in the model.

### 3.2.1. Workload as measured by ISA

Table 4 shows results of the regression analysis for ISA measures. The model using the three predictors is significantly better than just using the mean ISA score alone ( $F(3,28)=5.1$ ,  $p < 0.01$ ).  $R^2$  (adjusted) is 0.28. As such, the model predicts a modest proportion of variance in ISA score. T-tests show that all predictors contribute significantly to the model. Table 4 shows details of the model and associated predictors.

**Table 4: ISA multiple linear regression analysis (N = 32)<sup>1</sup>**

Predictor	B co-efficient (SE)	Standardised $\beta$ Co-efficient	t(28), p
HR Norm	3.38 (1.23)	0.42	2.74, $p < 0.05$
SDNN	-0.10 (0.04)	-0.32	2.08, $p < 0.05$
EyeNorm	4.22 (1.84)	0.35	2.30, $p < 0.05$

<sup>1</sup> Dependant variable: ISA measure of the workload, explanatory variables normalised Heart Rate, SDNN and normalised eye radius. ( $F(3,28)=5.1$ ,  $p < 0.00609$ , Adjusted  $R^2=0.284$ )

N=32	b*	Err-Type de b*	b	Err-Type de b	t(28)	valeur p
OrdOrig.			-4,88275	2,461410	-1,98372	0,057170
HR Norm	0,420089	0,153494	3,37665	1,233777	2,73684	0,010652
SDNN	-0,315906	0,152084	-0,09778	0,047075	-2,07718	0,047074
Eye Norm	0,352623	0,153398	4,21967	1,835639	2,29875	0,029193

As the adjusted  $R^2$  is 0.284, the following relationship explains around 28% of the variation of the workload as measured by ISA factor. As a result, we have:

- $ISA = -4.88 + (3.37665 \times HR \text{ Norm}) + (-0.09778 \times SDNN) + (4.21967 \times Eye \text{ Norm})$

### 3.2.2. Workload as measured by NASA-TLX

The same analysis is conducted to evaluate the relationship between NASA-TLX scores and the three physiological factors. Results are displayed in Table 5, which indicates that the whole relationship is significant ( $p < 0.00854$ ) and but that only two physiological factors (Normalised HR and Normalised Eye Radius) contributes significantly to the correlation.

**Table 5: NASA-TLX multiple linear regression analysis (N=32)<sup>2</sup>**

Predictor	B co-efficient (SE)	Standardised $\beta$ Co-efficient	t(28), p
HR Norm	26.20 (9.63)	0.42	2.72, $p < 0.02$
SDNN	-0.28 (0.04)	-0.12	0.76, $p > 0.05$
Eye Norm	40.98 (14.33)	0.44	2.86, $p < 0.01$

N=32	b*	Err-Type de b*	b	Err-Type de b	t(28)	valeur p
OrdOrig.			-60,7000	19,21543	-3,15892	0,003777
HR Norm	0,422956	0,155470	26,2030	9,63169	2,72050	0,011076
SDNN	-0,117149	0,154042	-0,2795	0,36750	-0,76050	0,453311
Eye Norm	0,444290	0,155373	40,9774	14,33024	2,85951	0,007928

Also, as SDNN is not a significant factor ( $t(28) = 0.76$ ,  $p > 0.05$ ), the regression analysis is conducted once again, but without this predictor. Results are given by Table 6. The regression is significant ( $p < 0.01$ ) and explains 28.0% of the variation of the workload as measured by NASA-TLX (adjusted  $R^2 = 0.28$ ), a small increase on the original model.

<sup>2</sup> Dependant variable: NASA-TLX measure of the workload, explanatory variables normalised Heart Rate, SDNN and normalised eye radius. ( $F(3,28) = 4.7357$ ,  $p < 0.00854$ , Adjusted  $R^2 = 0.266$ )

**Table 6: NASA-TLX multiple linear regression analysis without SDNN (N=32)<sup>3</sup>**

Predictor	B co-efficient (SE)	Standardised $\beta$ Co-efficient	t(29), p
HR Norm	25.93 (9.56)	0.42	2.71, p<0.02
Eye Norm	41.08 (14.22)	0.45	2.89, p<0.01

N=32	b*	Err-Type de b*	b	Err-Type de b	t(29)	valeur p
OrdOrig.			-61,4953	19,04696	-3,22861	0,003084
HR Norm	0,418639	0,154233	25,9355	9,55504	2,71433	0,011064
Eye Norm	0,445352	0,154233	41,0754	14,22511	2,88753	0,007266

As a result, we have the following relationship:

- $NASA-TLX = -61.4953 + (25.9355 \times HR \text{ Norm}) + (41.0754 \times Eye \text{ Norm})$

### 3.2.3. Stress measured by SACL

The regression analysis conducted for SACL measures with the three predictor variables shows that the SDNN factor is not significant ( $t(28)=0.49$ ,  $p>0.05$ ). Also the regression analysis is done with only the two other factors, as displayed by Table 7.

**Table 7: SACL multiple linear regression analysis (N=32)<sup>4</sup>**

Predictor	B co-efficient (SE)	Standardised $\beta$ Co-efficient	t(29), p
HR Norm	28.93 (10.84)	0.42	2.67, p<0.02
Eye Norm	44.24 (16.14)	0.43	2.74, p<0.05

Synthèse de la Régression; Variable Dép. : SACL (Feuil10 dans test.stw)						
R= ,55457946 R²= ,30755838 R² Ajusté = ,25980378						
F(2,29)=6,4404 p<,00485 Err-Type de l'Estim.: 4,6147						
N=32	b*	Err-Type de b*	b	Err-Type de b	t(29)	valeur p
OrdOrig.			-68,7472	21,61301	-3,18083	0,003485
HR Norm	0,416140	0,155971	28,9280	10,84232	2,66806	0,012356
Eye Norm	0,427501	0,155971	44,2424	16,14154	2,74090	0,010380

<sup>3</sup> Dependant variable: NASA-TLX measure of the workload, explanatory variables normalised Heart Rate and normalised eye radius. ( $F(2,29)=6.9150$ ,  $p<0.00350$ , Adjusted  $R^2=0.2762$ )

<sup>4</sup> Dependant variable: SACL measure of the stress level, explanatory variables normalised Heart Rate and normalised eye radius. ( $F(2,29)=6.4404$ ,  $p<0.00485$ , Adjusted  $R^2=0.2598$ )

Also, the relationship is significant and the two physiological values explain around 26% of the variations of the SACL values. We have the following relationship:

- $SACL = -68.7472 + (28.9280 \times HR \text{ Norm}) + (44.2424 \times Eye \text{ Norm})$

### 3.2.4. Situation Awareness measured by SART

With respect to Situation Awareness, only the HR Norm predictor was significant, ( $t(28)=0.04$ ,  $p>0.05$ ) for SDNN and ( $t(28)=1.46$ ,  $p>0.05$ ) for normalised eye radius parameters. Table 8 gives results for the regression analysis with the single remaining factor.

**Table 8: SART linear regression analysis (N=32)<sup>5</sup>**

Predictor	B co-efficient (SE)	Standardised $\beta$ Co-efficient	t(30), p
HR Norm	-42.22 (17.44)	-0.40	2.42, $p<0.05$

Synthèse de la Régression; Variable Dép. : SART (Feuil10 dans test.stw) R= ,40416624 R²= ,16335035 R² Ajusté = ,13546203 F(1,30)=5,8573 p<,02178 Err-Type de l'Estim.: 7,4942						
N=32	b*	Err-Type de b*	b	Err-Type de b	t(30)	valeur p
OrdOrig.			31,6978	18,54198	1,70952	0,097681
HR Norm	-0,404166	0,166998	-42,2185	17,44432	-2,42019	0,021778

Also, the relationship is still significant but heart rate values explain only 13% of the variations of the SART values. We have the following relationship:

- $SART=31.6978 + (-42.2185 \times HR \text{ Norm})$

### 3.2.5. Conclusions

Regression analyses between HPE factors and physiological data highlight significant relationships but the change in the physiological data explains only a small part of the variation of the HPE factors. Two measures of workload were used for Scenario 1 (ISA and NASA-TLX). Both the regression analyses show

<sup>5</sup> Dependant variable: SART measure of the situation awareness level, explanatory variable normalised Heart Rate. ( $F(1,30)=5.8573$ ,  $p<0.02178$ , Adjusted  $R^2=0.1354$ )



significant predictive relationships with physiological data and both relationships explain a little less than a third of the variation of the workload level. As the NASA-TLX is a prevalent measure of workload in the literature, we will keep only this measure of workload for the following regression analysis.

Also changes of three HPE factors are partly explained by two physiological measures (normalised Heart Rate and normalised pupil radius) with the following relationship:

$$F(2,29)=6.915, p=0.003$$

$$T_{HR}(1)=2.714, p=0.003$$

$$T_{EYE}(1)=2.887, p=0.007$$

$$\text{Adjusted } R^2=0.2762$$

- **WL (NASA-TLX) = - 61.495 + (25.935 × HR) + (41.075 × EYE)**

$$F(2,29)=6.440, p=0.004$$

$$T_{HR}(1)=2.668, p=0.012$$

$$T_{EYE}(1)=2.740, p=0.010$$

$$\text{Adjusted } R^2=0.2598$$

- **ST (SACL) = - 68.747 + (28.928 × HR) + (44.242 × EYE)**

$$F(1,30)=5.857, p=0.021$$

$$T_{HR}(1)=-2.420, p=0.021$$

$$\text{Adjusted } R^2=0.1354$$

- **SA (SART) = 31.697 + (-42.2185 × HR)**

It should be noted that these analyses rely on the global levels of workload, stress and situation awareness from the top of descent to the decision altitude. The analysis cannot reflect or predict sudden or short changes in the level of these parameters. Moreover, even if the physiological data have been normalised, their relationship with HPE factors are certainly partly subject dependant. Also for a more precise prediction of global HPE factors based on the use of physiological data, regression analysis could be done by pilot (but other data would be required). Finally, the study of the links between changes of HPE factors for small duration (within a scenario) and changes in physiological data has not been addressed here. It would require a more continuous evaluation of stress level and situation awareness.

### 3.3. Correlating performance and physiological data

As final step, the results from the previous correlation tasks are combined in order to use physiological measures as a predictor for performance. The results of this final correlation will be then applied and

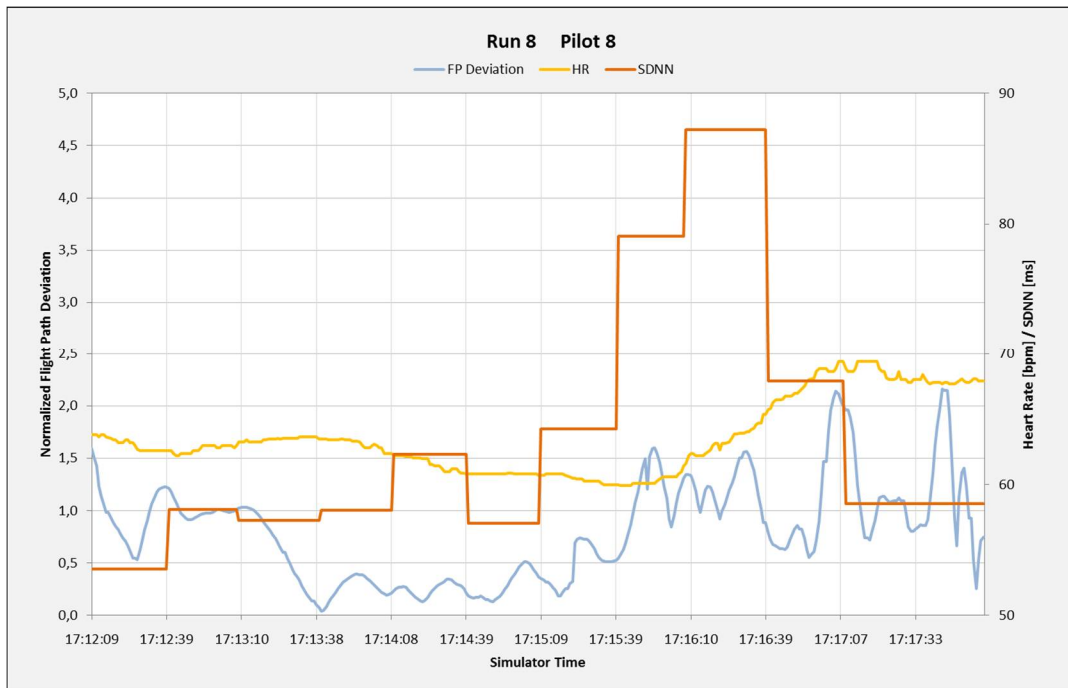
validated with Scenario 2, to see if the formula can be used to predict the degraded performances on this scenario, with different tasks and different performance measures.

This formula is a numerical integration of the formulas from the preceding two correlation tasks. The Stress, Workload and Situation Awareness terms have been replaced with their respective physiological relations. The same is done for the combo factor, but is shorthanded in the example below.

$Performance_{Predicted} =$

$$\begin{aligned}
 &C_{ST}(C_{ST_A}Phyio_A + C_{ST_B}Phyio_B + C_{ST_C}Phyio_C + \dots + C_{ST_N}Phyio_N) + \\
 &C_{WL}(C_{WL_A}Phyio_A + C_{WL_B}Phyio_B + C_{WL_C}Phyio_C + \dots + C_{WL_N}Phyio_N) + \\
 &C_{SA}(C_{SA_A}Phyio_A + C_{SA_B}Phyio_B + C_{SA_C}Phyio_C + \dots + C_{SA_N}Phyio_N) + \\
 &C_{Combo}(\dots)(\dots)(\dots) + \epsilon
 \end{aligned}$$

To facilitate the comparison between performance and physiological data, LOC, GS and SPD deviations together have been merged to get only one performance measure (as reported in Figure 52).



**Figure 52: Correlation between performance and physiological data**

The relation between physiological data and performance data is mediated by using the HPE factors. Mathematically, this means merging:

- HPE-Performance equation (see Section 3.1)
- with HPE-physiological expressions, i.e. one equation for each HPE factor (see Section 3.2).



This is simply using the HPE-Performance as the “parent” equation, and replacing the HPE factors with respective expressions of HPE factors in terms of physiological data.

The result of this combination of expressions is:

- **Performance =  $0.445 \times \text{HR} + 0.278 \times \text{EYE} + (0.130 \times \text{HR} + 0.139 \times \text{EYE}) \times (0.130 \times \text{HR} + 0.133 \times \text{EYE}) \times (-0.126 \times \text{HR})$**

The Performance equation can now be applied to Scenario 2.

### 3.4. Validating HPE concept with Scenario 2

The goal of this last step of the correlation exercise is to take the implicit equation from task 3 and apply it to Scenario 2 data. Using the physiological data from Scenario 2, a predicted performance is derived, which is compared to the competency performance ratings (see Sections 4.2.2 and 4.2.3). The better the predicted performance and competency performance align, the more universally applicable is the performance prediction model derived via the HPE using Scenario 1.

As Scenario 2 is a long-duration scenario with natural, less regulated HPE states, this exercise aims to conclude that A) the HPE concept applies in such a more natural scenario and B) the physiological measures are able to, via the HPE, assess performance. This is highly valuable because it would remove the need for measuring interim values such as workload, stress and SA, which still using their mechanisms.

Unfortunately, as not all the data is complete and reliable, this fourth correlation analysis relies only on the data for Pilot 5 and 10. This is because the competency data is only reliable for Pilots 3, 4, 5 and 10 (as these have multiple raters), and from these four sessions only Pilots 5 and 10 also featured complete HR and pupil-diameter datasets.

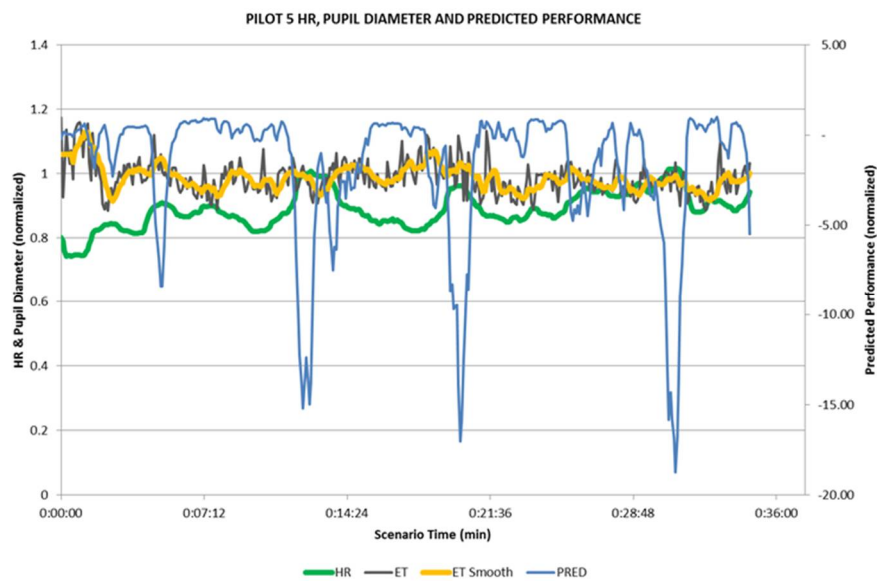
The analysis will compare the predicted performance using the equation from correlation task 3 with the three competency performance ratings by two methods. The first is a direct correlation analysis in which the  $R^2$  values are calculated for the three correlations between predicted performance and the respective competencies. The second analysis is a regression analysis in which the predicted performance is attempted to be explained by the three competency metrics (i.e. Situation Awareness, Decision Making and Application of Procedures, considered as independent factors), as illustrated in the equation below. The argumentation for this regression analysis is that the predicted performance measure represents total performance, and as such should be compared to the *total* set of competencies, despite the difference in the nature of these performances.

$$Performance_{predicted} = C_{CompSA}SA + C_{CompDM}DM + C_{CompAP}AP + \epsilon$$

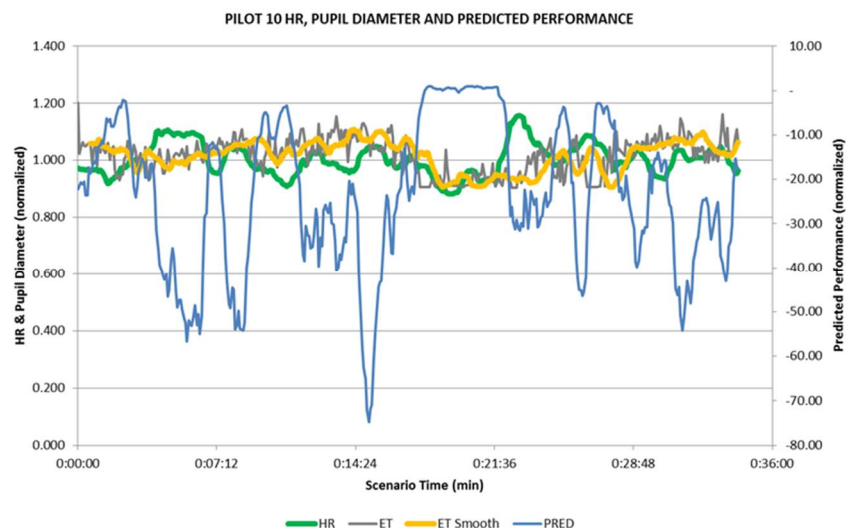
#### 3.4.1. From physiological data to predicted performance

The first step is to create the predicted performance measure. In order to do this the Scenario 2 physiological data (HR and pupil diameter) required some processing. The first step is to normalise these

physiological data streams against the same pilot-specific normal values used by ONERA in the previous correlation tasks, as normalised values were used in correlation task 2. Subsequently, the data has been re-discretised to a 5 second data stream in order to match the data frequency for the competency data. The last processing step pertained to eye-tracking only, and involved a smoothing function. As the eye-tracking data was sampled at 30 hertz, is provided quite jittery data even at 5 second resolution. Hence a smoothing function using a moving average of 30 seconds was applied to remove some of the hysteresis in the data. After this the data is applied to the equation that DLR derived in correlation task 3.



**Figure 53: Pilot 5 Heart Rate (HR), Pupil Diameter (ET) and Predicted Performance (PRED)**



**Figure 54: Pilot 10 Heart Rate (HR), Pupil Diameter (ET) and Predicted Performance (PRED)**

### 3.4.2. Comparing predicted and actual performance

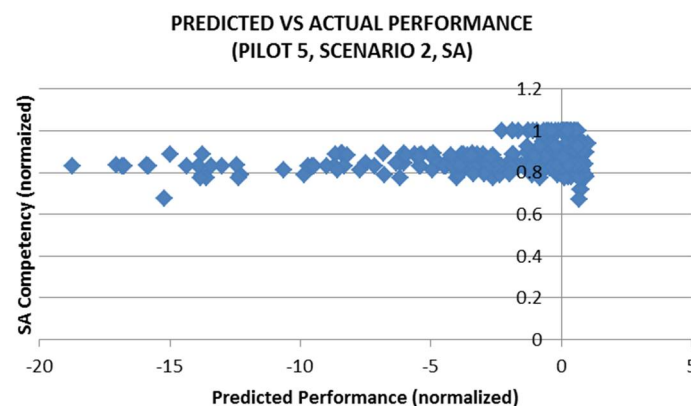
In order to make any conclusions about the predicted performance derived, it must be compared to other performance metrics and analysed for alignment/correlation. The analysis will only be performed for pilot 5 and pilot 10 (due to data limitations), and will perform four analyses. The first three analyses will be to correlate the predicted performance with the different competency performance indications (SA, DM and AP) independently. The fourth analysis will be a regression analysis in which all three competencies are combined as independent factors.

The table below shows the  $R^2$  values of the initial correlation analysis for Pilot 5. The values in the first column indicate the correlation between the predicted performance (PRED) and the three competencies independently (SA, DM, APP). Representing only 3%, 2% and 0.07%, these correlations are no-existent. On a side note, there is a peculiarly high correlation between SA and DM (0.422), although this is still not a strong correlation.

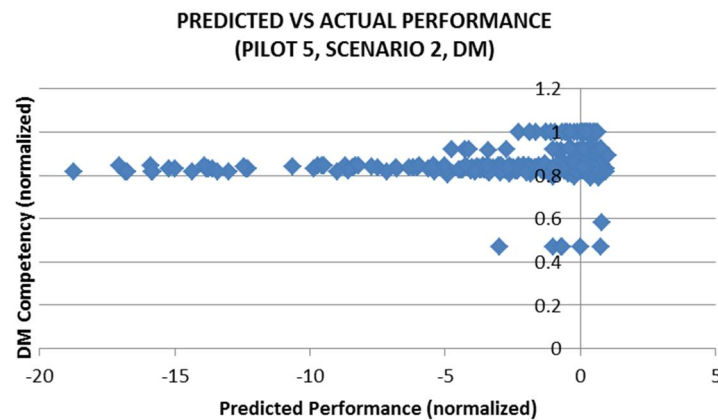
**Table 9: Pilot 5 performance correlation analysis**

	<i>PRED</i>	<i>SA</i>	<i>DM</i>	<i>APP</i>
<i>PRED</i>	1			
<i>SA</i>	0.03357	1		
<i>DM</i>	0.016729	0.422606	1	
<i>APP</i>	0.000798	0.136732	0.060777	1

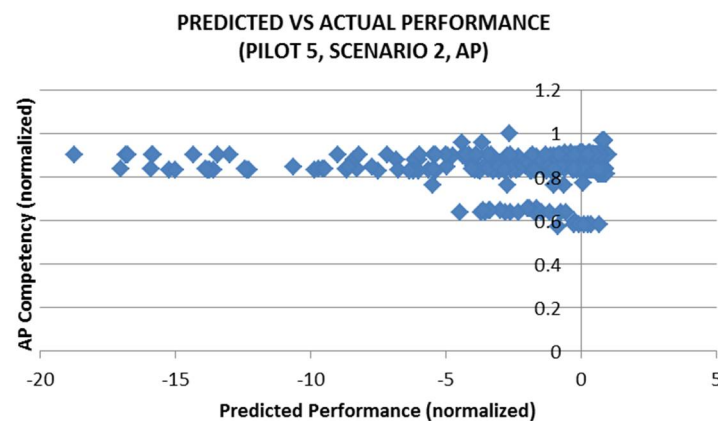
A further regression analysis analysed if the three competencies could collectively relate to the predicted performance (a total-performance indicator), however this analysis featured an  $R^2$  of 0.0356, which confirms the low correlation between the predicted performance and competencies, for the situation of Pilot 5. To illustrate these analyses, the three figures below show how the predicted and competency performance metrics look if plotted against each other. Clearly the horizontal variation in predicted performance isn't reflected in a vertical variation of the competencies, indicating a lack of relation.



**Figure 55: Pilot 5 performance correlation analysis (PRED-SA)**



**Figure 56: Pilot 5 performance correlation analysis (PRED-DM)**

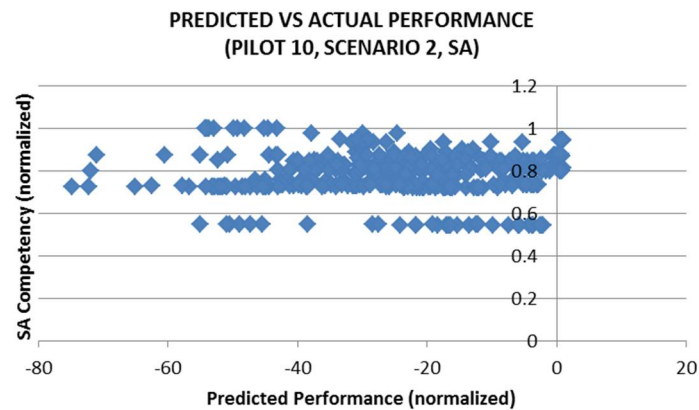


**Figure 57: Pilot 5 performance correlation analysis (PRED-AP)**

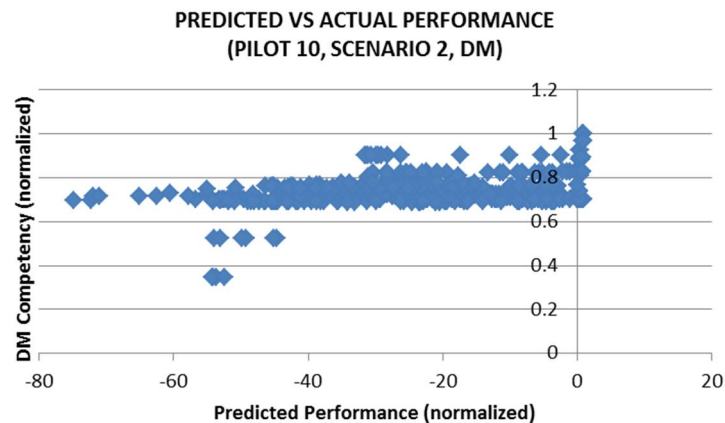
The same analysis is performed with Pilot 10. The table below shows the  $R^2$  values of the correlations made, and also indicate a low correlation. The most prominent correlation is that of the predicted performance with decision making, yet this only accounts for explaining 18% of the variance. The regression analysis of the combined factors produced an  $R^2$  of 0.218, marginally better than the DM  $R^2$  alone. The figures below also visualize the plotting of predicted performance against the three performance metrics.

**Table 10: Pilot 10 performance correlation analysis**

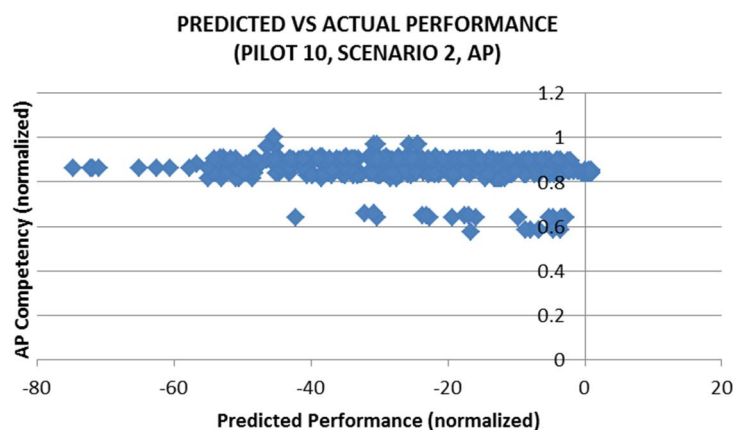
	<i>PRED</i>	<i>SA</i>	<i>DM</i>	<i>APP</i>
<b>PRED</b>	1			
<b>SA</b>	0.004826	1		
<b>DM</b>	0.186205	0.03272	1	
<b>APP</b>	0.037608	0.004402	0.00152	1



**Figure 58: Pilot 10 performance correlation analysis (PRED-SA)**



**Figure 59: Pilot 10 performance correlation analysis (PRED-DM)**



**Figure 60: Pilot 10 performance correlation analysis (PRED-AP)**

The results from the above analyses for Pilot 5 and Pilot 10 indicate a very weak relation at best between predicted and competency performance, indicating that the variations in the competency performance are not sufficiently explained or mirrored by a change in the predicted performance. Even when

comparing the predicted performance to a the total-performance regression using all three competency metrics, the relations remain weak.

### 3.4.3. Conclusions on the applicability of the HPE model

The low correlation between the predicted and actual performance measures leads to the conclusion that the mathematical construct for proving the HPE could not successfully translate from Scenario 1 to Scenario 2. There are multiple possible explanations for this invalidation. First and foremost, the performance metric in Scenario 1 (flight path deviation) and the performance metric in Scenario 2 (competency measures) are quite different in what they observe, and as such do not necessarily correlate. Secondly, the prediction formula is a mathematical construct designed around the multi-factor HPE concept, but also permits accumulation of errors. As such the model is possibly prone to sensitivity, which is somewhat visible in the predicted performance dataset, which has several major peaks and valleys. Third and lastly, the scoped HPE concept using only three factors (workload, stress & SA) may not cover all the facets of performance, and therefore be limited in its predictive power. For this reason the research project must conclude that the HPE concept cannot be validated using the previous correlation analysis steps. In the event that a correlation were to be found, the small amount of data used for this validation exercise also restricts the conclusive power of this validation exercise. Nonetheless, this exercise does provide a framework for future HPE validation exercises.

### 3.4.4. From actual performance to physiological data

As an alternative to the formulaic construct derived in the third correlation task and (in)validated in the previous section, it may be possible to infer a direct correlation between Scenario 2 physiological data and the competency performance metrics. Although such a correlation exercise may reveal a useful predictor for competency performances, it does not justify or validate the HPE concept as it becomes a numerical exercise without the theoretical construct of a multi-factor human model.

However, this does not discount the value of such an explorative question, and it could be a valuable alternative finding than the HPE model used from the onset. In order to make a reliable (explorative) model, there must be many sets of data to be correlated to verify any significant link. As both physiological and competency data will be collected for the Scenario 2 in both the Braunschweig and Thales experiments, it would be beneficial to perform this analysis using the collective dataset of all setups. Hence another attempt at correlating performance with physiological measures will be made in work package 6.4.

## 4 PRINCIPLES AND CONSIDERATIONS FOR HMI DESIGN TO SUPPORT RECOVERY FROM PERFORMANCE DEGRADATION

Despite the validation tasks didn't give the expected results, and the performance decrement cannot be directly associated to each specific factor under investigation, or to a variation in the pilot's physiological status, the expert analysis of the data collected during the simulations in Braunschweig and the subjective performance assessment provided useful hints for the HMI re-design and suggestions for performance recovery.

This section is dedicated to the deep dive analyses of pilots' performance in the two scenarios for the identification of the contextual conditions and factors that led or contributed to degraded performance. The result of these analyses will be used to develop suggestions for HMI improvements and other measures to support the performance recovery.

### 4.1. Support the recovery of Pilot Flying performance

The identification of Scenario 1 critical performance points and areas/strategies for recovery is based on the analysis of pilot's self-assessed performance through the performance curves (for more details, see D6.3), combined with the explanations collected in the debriefing phase. For each run, the pilots' position onto the curve and the identification of the points with performance decrement allowed the identification of critical areas where HMI improvements or new tools, systems or features could have helped the pilots to face the situations encountered during the simulation.

On the basis of pilot's debriefings, some recurrent issues were identified and changes or improvements for existing A320 on-board systems and interfaces were discussed with the subjects. Specifically, three areas for improvements emerged:

- **Electronic Flight Bag (EFB):** the interaction with EFB should be simplified by means of a better information architecture (to facilitate information search) and tactic feedback, at least for the more common functions. These changes can improve pilot's performance in time critical situations (such as the final approach segment) when pilot cannot spend time looking at EFB, as his/her attention is needed elsewhere.
- The **Navigation Display** HMI resulted too cluttered, it should be simplified, and the pieces of information displayed could be reduced.
- Integration of **wind information** into the **Primary Flight Display (PFD)** can help to have wind in the scan path and facilitate the calculation of the correction angle. PFD could also integrate **Track indication / visualisation of the optimal descent profile** compared to the actual aircraft profile (Embraer-like), and provide a warning if the aircraft is diverting from the normal trajectory. This could be really useful especially during non-precision approach or in case of strong cross-wind. Additionally, **thrust lever** could be reported nearby speed indicator to facilitate the correlation



between speed and power settings and faster the cross-check of aircraft parameters, and **speed brake** information should be moved from the side of the cockpit to the front display.

Additionally to these changes, some pilots wished for the implementation of a **Head Up Display** (HUD) system showing a set of relevant aircraft parameters such as speed, altitude, glide slope, flap settings, and wind. The immediate access to these pieces of information can be particularly useful in the final approach phase, when pilots need to look at the runway and the on-board information scan is consequently reduced as much as possible. In low visibility conditions, HUD can also show a picture of the runway to improve pilot's situation awareness.

Other potential improvements emerged during the debriefings concerned the **on-board visualisation of ground-related information**, particularly real-time updates on the runway status (wind near the runway, runway conditions etc.), and **visual information on terrain situation**, likewise google maps. As terrain is one of the flight main risks, visual information on terrain could work as a back-up solution in case of localiser failure; on this map, obstacles can be shown together with the safe areas terrain-wise. A **display showing the aircraft clearance to land** could help pilots as well, in particular in high workload conditions when the radio communications are more likely to be skipped or forgotten. A written message (following the radio communication) can stand there and be read again by the pilot if he/she doesn't recall the voice message.

The simulated experience of fuel shortage landing raised the need for well-timed **visual or audio warnings** drawing pilot's attention on the **remaining fuel**. Even more, some pilots envisioned an on-board decision support system able to provide an estimation of the fuel consumption depending on the different choices of trajectory (go-around, delay vectors etc.). Another decision support system imagined was a **system able to support performance calculation** by correlating aircraft parameters/configuration with environmental information, and inform the pilot if something goes in the wrong direction and performance limits are being approached (is the wind becoming too strong? Is the rate of descent too strong? Is the terrain becoming too close?). For example, this system can be able to predict an unstable approach due to tailwind and inform the pilot that he/she has to change the power setting to avoid this situation. Turbulences could be predicted as well, as the patterns that lead to turbulences are known but are difficult to be recognized by pilots, as they depend on the correlation between climbing rate and temperature changes. Also, another subject mentioned the opportunity to have a **support for information prioritization**, which could help pilot to recover from attentional tunnelling.

It can be noted that, among the different HMI communication channels (visual, audio, tactile), pilots show a strong preference towards visual channel for non-critical communications or "kind warnings" (far before the situation becomes dangerous), while the audio channel should be limited to the critical warnings that require an immediate intervention. Also, almost all the inputs provided went in the direction of short-term HMI improvements or systems implementation, while no out-of-the-box ideas were mentioned by pilots during the debriefings.



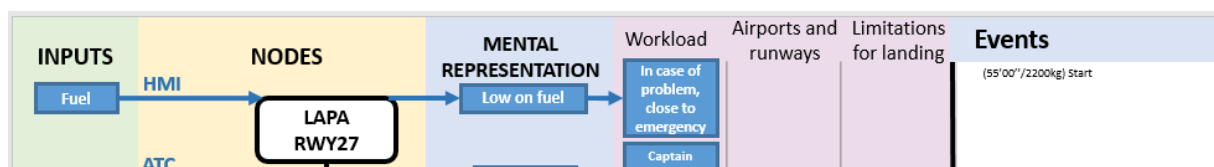
## 4.2. Support the recovery of Pilot Monitoring performance

### 4.2.1. Pilots Mental Representations

Due to Scenario 2 level of complexity, to understand the different pilot choices and behaviours during the execution of the scenario, the results of the cognitive walkthrough conducted by CATIE during the simulations were used to construct the pilots' mental representation and to define its impact over 3 parameters: workload; airport and runways selection; and limitations for landing.

To facilitate the comparison between pilots' actions and the "expected behaviour", the analysis was structured by distinguishing the following categories (see Figure 61):


- **Inputs:** Relevant information provided to the pilots by the HMI, the Air Traffic Controller (ATC) or the Pilot Flying (PF). The inputs are the cues that should create/change the mental representation. Each input has been linked with a legend that defines the source of information (PF - Pilot Flying; ATC - Air Traffic Controller; LAPA - Landing Calculation; OMB and QRH - Operational Manuals; HMI - Airplane instruments).
- **Nodes:** or "scenario phases". Moments of the scenario during which the pilot had to perform procedures or take decisions.
- **Mental Representation:** how the inputs and the situation were understood (meaning, impact, consequences, etc.).
- **Workload, Airports and runways, Limitations for landing:** Impact of the mental representation on each parameter. They will provide an approach about the impact of misinterpretations.
- **Events:** Timeline with the events observed during the run.

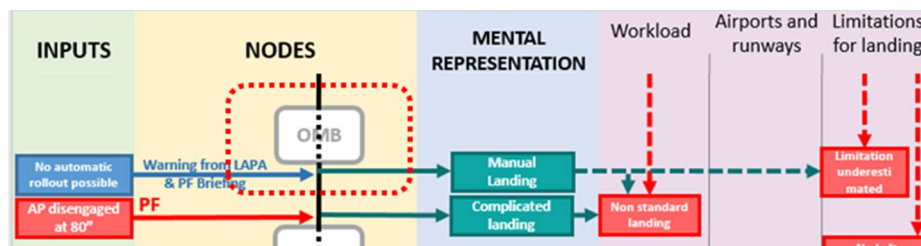


**Figure 61: Structure of Scenario 2 and mental representation of the pilot**

Specific colour coding has been used to indicate the level of performance of the pilot with respect to information collection (source and timing), information understanding and situation awareness.

- In the **Inputs** column:
  - **GREEN BOX** indicates that the PM behaved better than expected in terms of input searching, knowledge, briefings, etc.
  - **BLUE BOX** indicates that the PM perceived the inputs without any additional help.
  - **CYAN BOX** represents an acceptable input perception, but not at the best of PM's performance.
  - **RED BOX** Indicates that the PM missed the cue, and that the captain or the ATC gave him the correct information.

- In the **Mental Representation** column:
  - GREEN BOX** indicates that PM understood the situation better than expected, and was able to anticipate decisions.
  - BLUE BOX** Means that PM understood the input and situation without any additional help.
  - CYAN BOX** Means that PM understood the input and situation with some additional help.
  - RED BOX** Means that PM didn't understand the inputs correctly, and that he didn't have an acceptable representation of the situation. The box  means that the pilot missed an input.
- In the **Nodes** column, **GREY BOXES** indicate that decisions were not taken or specific procedures were not performed (see example in Figure 62)



**Figure 62: Nodes column - OMB Procedure not performed**

A complete overview of the application of MEntal Representation Impact Analysis (MERIA) model is reported below, while an overview of the results of the analysis for all pilots can be found at the end of the section. Pilot 5 Mental Representation is presented in Figure 63.

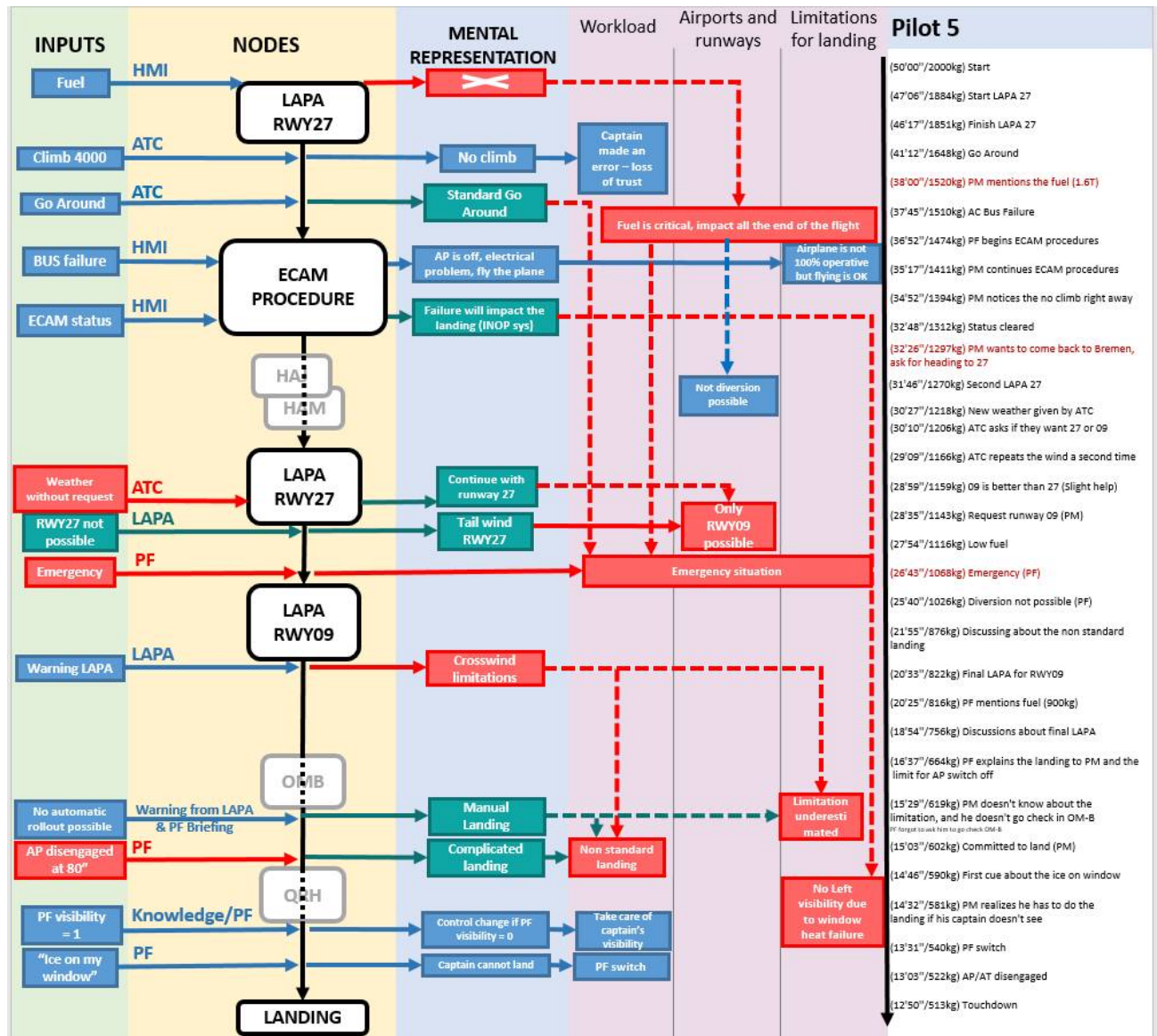
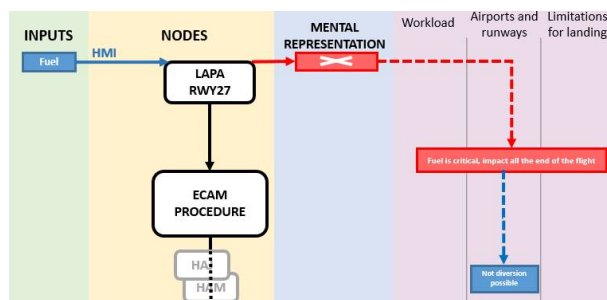


Figure 63: Mental representation of Pilot 5

The scenario starts with a low fuel condition.

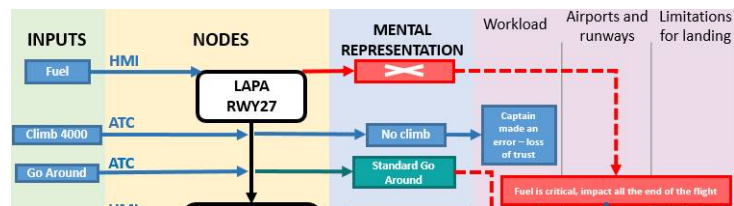
- Despite the **INPUT (Fuel)**, PM didn't realise about fuel status at the very beginning, thus in the Mental representation there is a box representing the missing input. The delay in realising the fuel stats had an influence on the PM workload, on his decisions on airport and runway, and thus in realising limitations for landing.



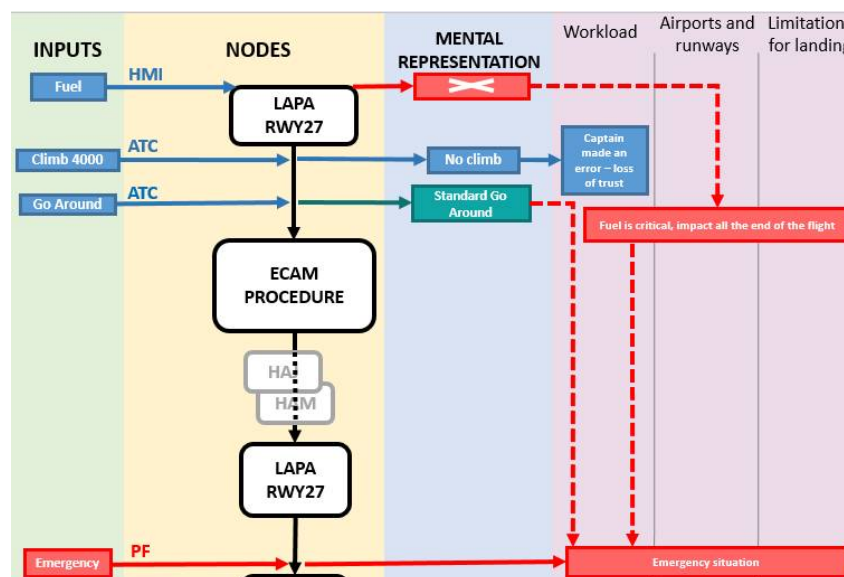
Basically, the pilot realised very late that his only landing option was Bremen and that no diversion was possible.

- The second **INPUT (Climb 4000)** given by the ATC required an action by the PF. In this case, PM realised that PF didn't pull the button to climb. Despite the recognition of the input, the missed action by PF had a consequence on PM workload, as a loss of trust in PF was reported and this put an extra amount of work for PM to check the actions of PF.

- A third **INPUT (GO AROUND)** given by the ATC was initially misinterpreted by PM, as it was seen as a standard procedure without critical considerations. Just a few minutes later the pilot



realised about the lack of fuel. The mental representation of the GO-AROUND (a standard procedure) without the perception of FUEL status shows an impact over the time where the EMERGENCY has been declared (Figure 64). Moreover, the emergency is really motivated by recommendations of the captain.



**Figure 64: Emergency is declared late**

- The fourth input - **INPUT (BUS failure)** – came from the HMI and was a critical failure that the PM understood correctly and properly handled. Moreover, the pilot rapidly understood that it was possible to fly and land despite of this problem.
- Soon after the BUS failure, another input came from the HMI: **INPUT (ECAM status)**. In this case, the interpretation of the failure was not as expected, despite it remained at an acceptable level. In particular, the PM didn't consider the importance of "L WNDW HEAT" failure. However, his future decisions showed that he had the knowledge to interpret the implications of low visibility

in these weather conditions. In fact, in the landing phase, when the PF reported low visibility in his window, PM reacted well to control the situation (PF switch). At this stage, Pilot 5 didn't consider new airports because he had clear in his mind that no diversions were possible due to fuel limitation.

- After the BUS failure, there was a wind shift making impossible to land by the RWY27 - **INPUT (New weather and LAPA results)**. When PM started to calculate the LAPA to land on runway 27, the ATC transmitted an update of weather conditions that made impossible to land on that runway. However, PM continued to perform the calculations for runway 27 using the new ATC data. He realised the limitations for RW27 (tail wind) only thanks to LAPA results. These LAPA results were rapidly understood and interpreted as "Only RWY09 possible".

Also, results of LAPA RWY09 showed a warning message: "Please check crosswind limitations on contaminated runways in OM-B. CWC limit for automatic rollout is exceeded. CWC limit for automatic approach is 20 kt." (Figure 65 and Figure 66 - **INPUT (Warning LAPA and debriefing with captain)**). The OM-B specifies that the CWC (Cross Wind Component) limitation implies a landing manoeuvre without Automatic Rollout. However, PM didn't understand the LAPA warning message and consequently he didn't check in the OM-B. From his discussion with PF, it came out that PM understood that "automatic rollout" was not possible, however he didn't know that together with that PF should have disengaged the autopilot at 80 feet to perform this landing.

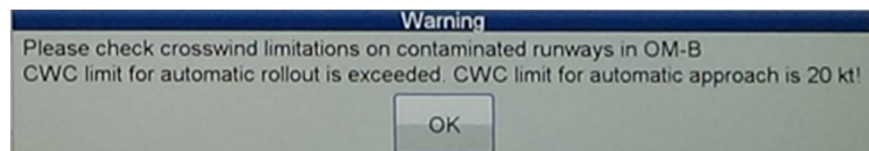


Figure 65: Warning message of LAPA RWY09

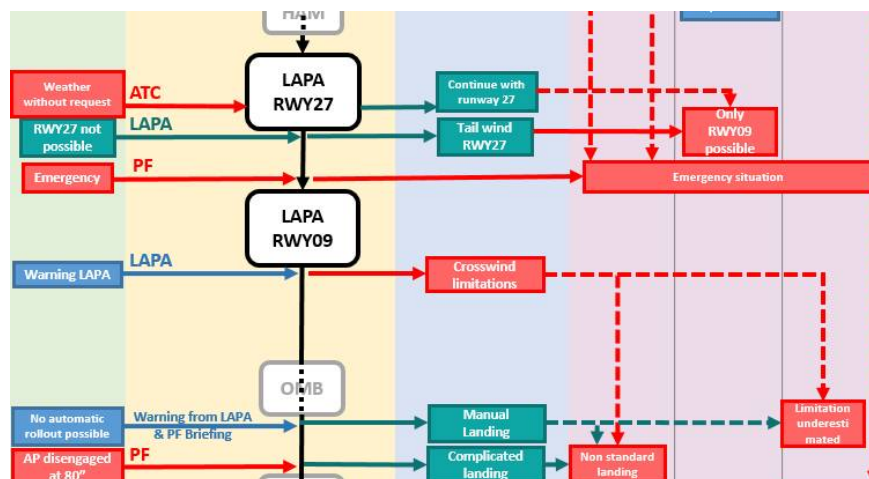


Figure 66: INPUT - LAPA results and Crosswind limitations

- Finally, during landing briefing PF reported low visibility from his window - **INPUT (PF visibility = 1)**. PM shown the necessary knowledge to understand the importance of PF low visibility. Thus,



he reacted quickly to that input when the captain informed him that there was no visibility in his window and went for a task switch to land.

The analysis of mental representations of all the 10 pilots performing Scenario 2 brought to light that the main differences among pilots' performance depended on the identification and interpretation of three key aspects: Fuel Status; Electrical failure; and Weather. Each aspect is presented in detail in the following sections.

#### **4.2.1.1. Fuel status**

Pilots can be divided into two groups on the basis of the perception time of the fuel status:

- Group 1 - Early detection: 40% of the participants realised the fuel situation between the beginning of the scenario and before the Go Around. This group of pilots had a better management of resources and spent less time in doing landing hypothesis. In fact, no other airports were considered, or the option was quickly discarded. Also, 75% of pilots in this group declared emergency, by themselves.
- Group 2 - Late detection: 60% of the participants realised the fuel situation after the GO Around. This implied that pilots in this group spent a lot of time considering different options, even unfeasible options. The analysis of this group of pilots showed that the average time spent on considering a new airport is four times longer than pilots in Group 1. Also, 70% of pilots in this group needed a cue from the captain to declare emergency.



**Figure 67: Fuel state in the HMI**

#### **4.2.1.2. Electrical failure**

In case of electrical failure, PM is expected to identify the name of the failure, understand the type of malfunction, the time needed to solve the problem and the implications this failure has on the plane and on the flight. For the simulated type of electrical failure, the expected reaction was a PF switch, with the subjects taking the control of the flight.

In managing the failure and identifying the right things to do, pilots were supported by the ECAM status, available on the Navigation Display during the ECAM procedure (Figure 68). However, the list of failures was not prioritised and a part of the listed systems may be lost due to the characteristics of the HMI (consideration available in the OMB and not in the HMI; see Figure 69).



Figure 68: Electrical failure situation



Figure 69: ECAM status of Bus Failure

The screenshot shows a flight simulation software interface for a missed approach procedure. A red circle highlights the 'Ldg Type' dropdown menu, which is set to 'MAN LDG'. The interface includes various input fields for aircraft performance, runway data, and a table of missed approach data.

Runway	Identifier	LDA	Slope	G/S	Max
09	EMERGENCY ONLY	2334	0.0	3.0	2.5
09	EMERGENCY ONLY A1087-16	2334	0.0	3.0	2.5
09		2034	0.0	3.0	2.5
09	A1087-16	2034	0.0	3.0	2.5

Act LAW: 64.0 t VPilot: 0 LOG CG: >= 25% Ldg Type: **MAN LDG** RCAM: Medium to High

Adt LAW / MLAW: 64.0 VLB: 134 MLAW: 64.5 PLAW(I): 67.8

CONF: FULL Manual 1923 2211 TWY D 1170  
APPR CLB Grad: LO 2017 2319 TWY E 1610  
5.8 % MED 1965 2259 TWY F 1990  
G/A CLB Grad (avg): 5.6 %  
LDG CLB Grad: 19.2 %

Field Length: 67.8  
APPR CLB: 75.8  
G/A CLB: 74.7  
LDG CLB: 82.3  
Brake Energy: 90.0  
Tire Speed: 90.0

BASED ON PUBLISHED MISSED APPROACH - 1 EO accel alt: 2100ft

**Figure 70: Considerations for LAPAs coming from electrical failure**

The analysis of pilots' behaviour showed that 30% of pilots didn't remember the considerations for landing from the "abnormal ECAM procedure" (Figure 70, Only CATII possible) during LAPAs. Thus, they were not able to correctly enter these considerations into the LAPA without the captain's help.

Even more, several misinterpretations were made by PMs, which affected their future decisions:

- The inoperative "Reverser 1" wasn't noticed or remembered by 30% of the pilots, meaning they might had an erroneous perception of the necessary landing distance.
- The "left window heater" failure, in combination with the bad weather, was only recognised by 20% of the pilots as a possible condition for PF switch, due to low visibility of the captain.
- Most of the pilots needed help from the captain or OMB to conclude that the "nose wheel steering" in combination with the crosswind, which could result in a runway excursion and that thus required a manual roll-out.

#### 4.2.1.3. Weather

Weather was a critical parameter in the Scenario 2, especially due to the impact of wind change on landing procedure. In fact, the wind change resulted in a change in the landing runway, and the PM was supposed to anticipate this. The results from the group analysis said that:

- After completing the ECAM procedure, only 30% of pilots thought about asking for the new weather. Nothing in the HMI indicated that the weather may had changed and the ATIS (Automatic Terminal Information Service) was mentioned only by one pilot. That means that 70% of the pilots, without external help, would have tried to land with a tailwind.
- Only 20% of the pilots understood the "warning message" for rollout with crosswind. The other 80% forgot the failure (Nose wheel steering) that already forced the Manual Roll-out (Figure 71).



- CATII landing was mandatory and in combination with information in the OM-B (or pilot's knowledge) implied that a manual landing (with 80" limit to disengage Auto Pilot) had to be done. Only 50% of the pilots knew this limitation, the rest of them had serious problems to find this information. Only 1 pilot found the correct section of the OM-B (Figure 72).
- Even after the pilots got the new weather (with tail wind), 60% didn't realise that landing in RWY27 was not possible. Nothing in the LAPA helped them in understanding this immediately (they must wait the calculations).

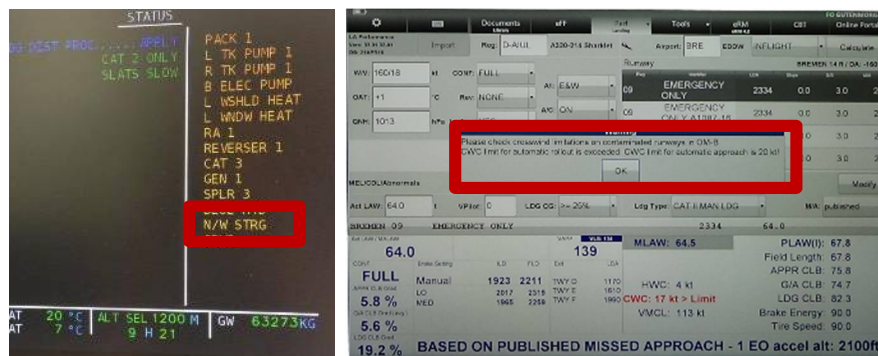


Figure 71: Two different HMI messages meaning that Roll-out must be in manual mode

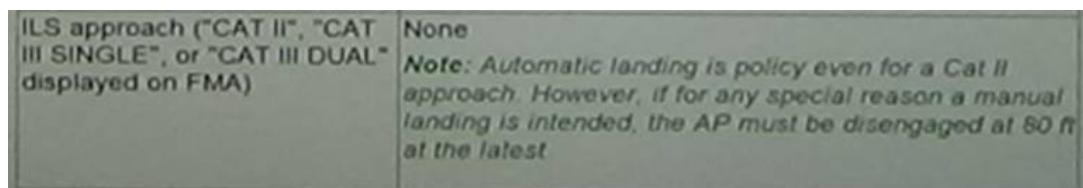


Figure 72: Limitations for landing in OM-B; CATII procedure.

#### 4.2.2. Competence evaluation

Together with the analysis of Pilot's Mental Representation, the performance of the Pilot Monitoring in Scenario 2 was assessed by a group of observers through a tool supporting the continuous rating of the following core competences:

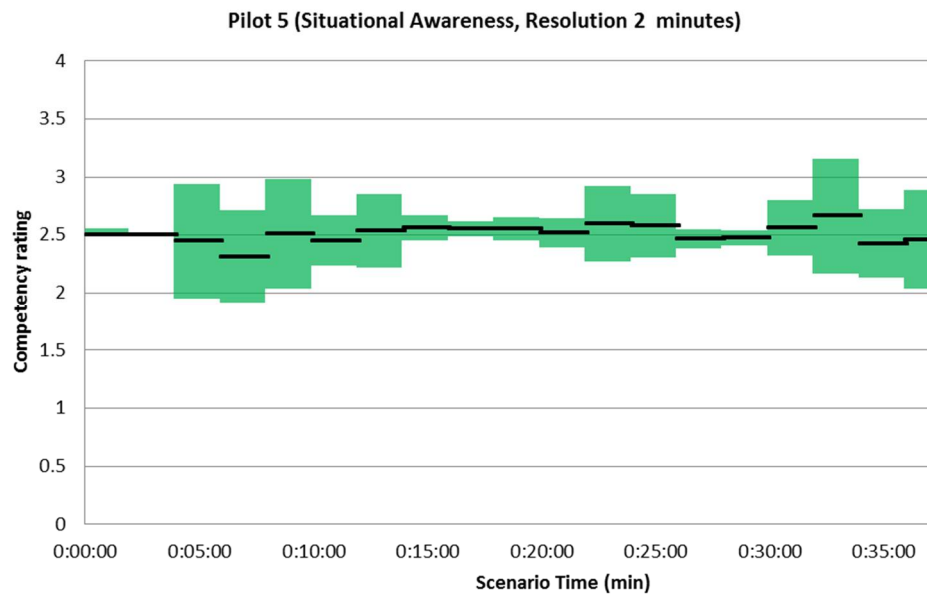
- Situation Awareness (SA);
- Problems Solving and Decision Making (DM);
- Application of procedures (AP).

The full method can be found in D6.3. To sum up, a group of observers individually played the videos from the PM scenarios and meanwhile provided a rating for the pilot performance on the abovementioned three competencies at any given time. The observers were provided a scenario description and example behaviours per event and the corresponding rating, to standardise the rating frameworks and thereby increasing the interrater reliability. A four point scale was provided but the rating was set by means of a slider that allows any ratings in between two discrete values. Each change in performance rating was time

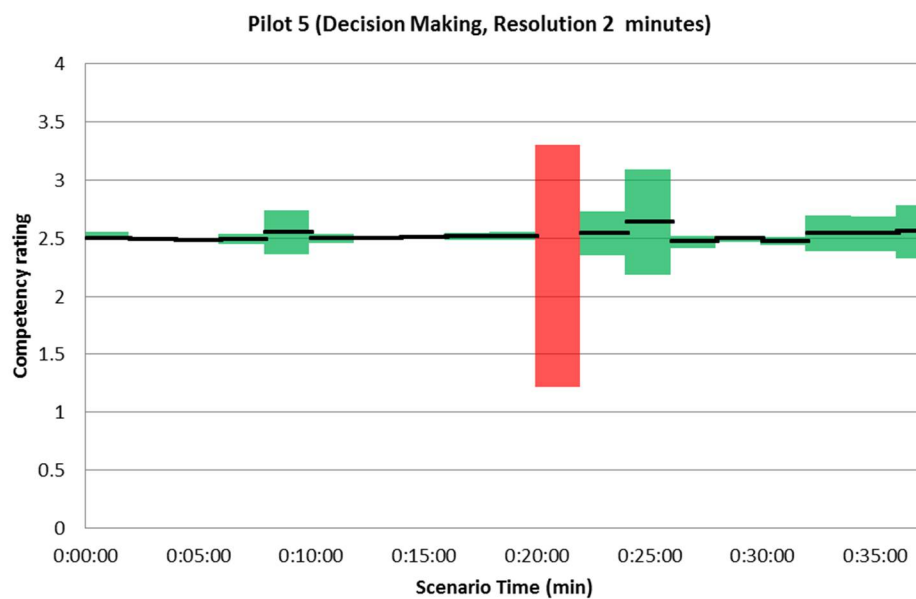
stamped and saved. As the competence assessment for Scenario 2 is based on subjective evaluations by experts, a measure of rater-agreement of the competency data – called Inter Class Correlations (ICC) – is provided. The ICC analysis investigated the rater-agreement for data resolutions of 5 seconds, 1 minute and 5 minutes, considering both block averages and moving averages. As the ICC changes with a change in resolution, it is possible to determine a minimum valid resolution for competency analysis. From a training perspective, this is useful information when developing methods to assess competence in realistic scenarios. From a research perspective this information is useful to define a crew performance metric for a realistic scenario, where more objective (or rather, flight performance driven) measures become more convoluted due to the complexity of the scenario, and therefore less powerful.

Several dimensions have been considered to augment ICC analysis (resolution variation, ICC between independent raters or with respect to an average score etc.). Among them, the Temporal Reliability Analysis (TRA) can be used to determine how reliable (i.e. aligned) the ratings are at any moment in time throughout the scenario. This may indicate events in which instructors readily align, or moments at which there is a difference of opinion. By identifying, understanding and addressing these moments of difference, the assessment of competencies can be improved. This analysis can provide a measure of temporal (un)certainty when comparing competency data to MERIA model results. As this particular TRA will be used for this research context, the resolution used will be the minimum resolution for an average-ICC agreement, as only the average rating data stream is required for a comparison with MERIA's dataset.

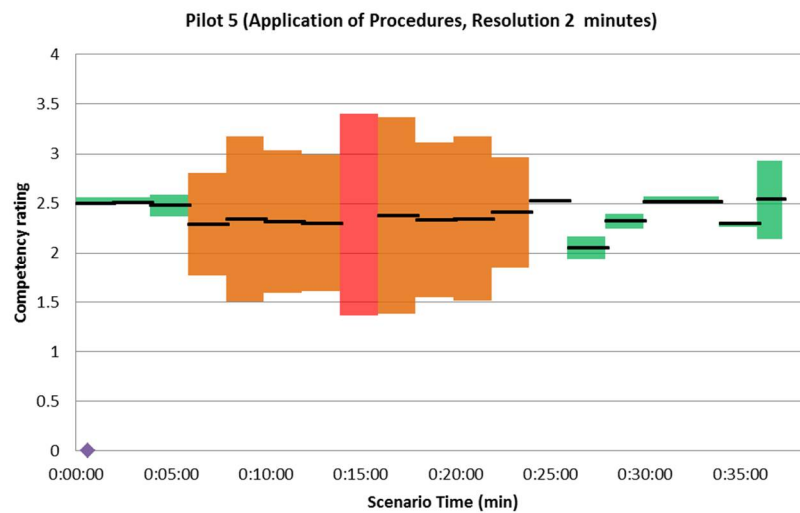
Each competency has its own TRA as depicted in the three figures below, at a resolution of 2 minutes. For each discretized moment, the standard deviation is calculated and depicted as a bar. Green bars indicate that the 90<sup>th</sup> percentile range is smaller than 1 competency point difference. An orange bar indicates a range between 1 and 2 point differences, and red bars indicate a 90<sup>th</sup> percentile range greater than 2 points. This provides a clear, visual representation of the (un)certainty of the competency ratings. The competency rating 4-points scale goes from 1 (Unacceptable) to 4 (Exceeds), with 2 and 3 indicating respectively indicating Below expectations and Meets expectations.



**Figure 73: TRA of Pilot 5 Situational Awareness**



**Figure 74: TRA of Pilot 5 Decision Making**

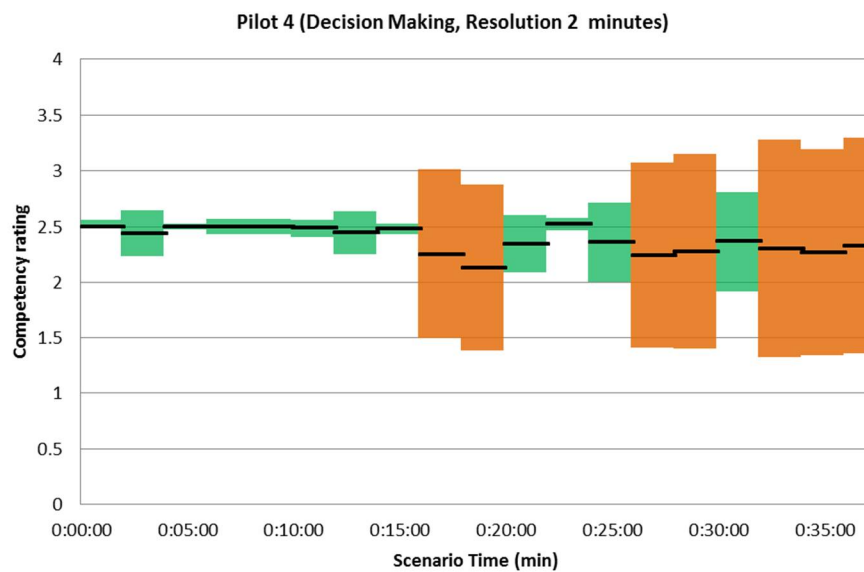


**Figure 75: TRA of Pilot 5 Application of Procedures**

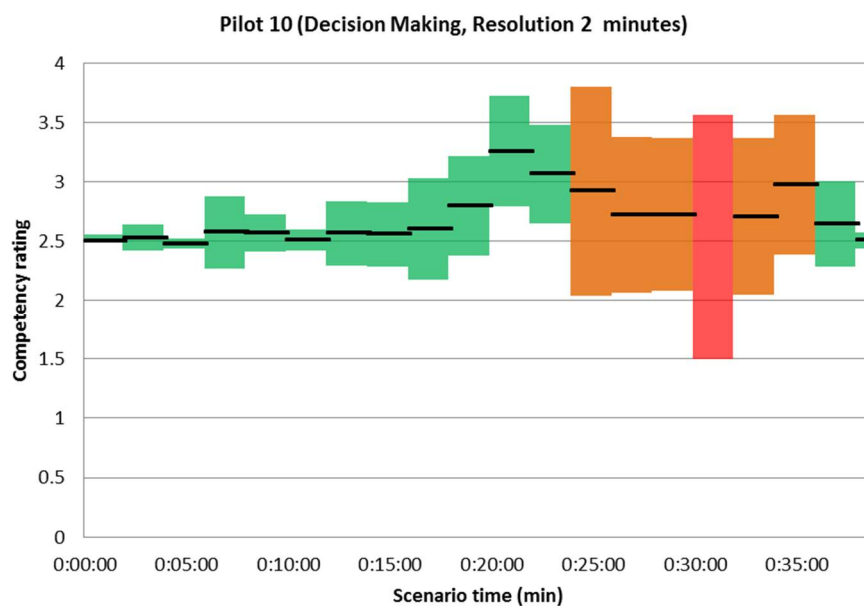
The above plots show three different competencies for the same crew. To contrast, below are the decision making plots of three other crews to contrast between crews.



**Figure 76: TRA of Pilot 3 Decision Making**



**Figure 77: TRA of Pilot 4 Decision Making**



**Figure 78: TRA of Pilot 10 Decision Making**

To improve the training of competencies, such local moments of high variance can be identified in the scenario video and further investigated by experienced instructors. Performing TRA's for all competencies and all pilot sessions (using individual ICC minimum resolutions) can provide a database of events which are prone to rating disagreement. These can be further analysed to be better supported in competency assessment, in an effort to reduce rater-disagreement. Such a broad analysis of the cause of variance has not yet been performed, but can be part of future work in this field.

#### 4.2.3. Results from competence evaluation and cognitive walkthrough matching

With the TRA's available, the competency data can be effectively compared against other FSS data. The competency data is a performance metric for Scenario 2 (where there are no other performance metrics available). Such a performance benchmark can be used to indicate, using other partners' data sets, which pilot states, behaviours and cognitive processes are present during good performance, and which during poor performance. This in turn can guide new training and cockpit interventions in the right direction.

One such comparison has been performed with MERIA model on mental representation of each pilot's supposed understanding of the situation as they are executing Scenario 2. This mental representation is in turn judged to be a good (constructive) or bad (misleading) representation. By understanding how both the good and bad representations develop, this analysis can be fundamental to designing better cognitive support in new HMI's and automation.

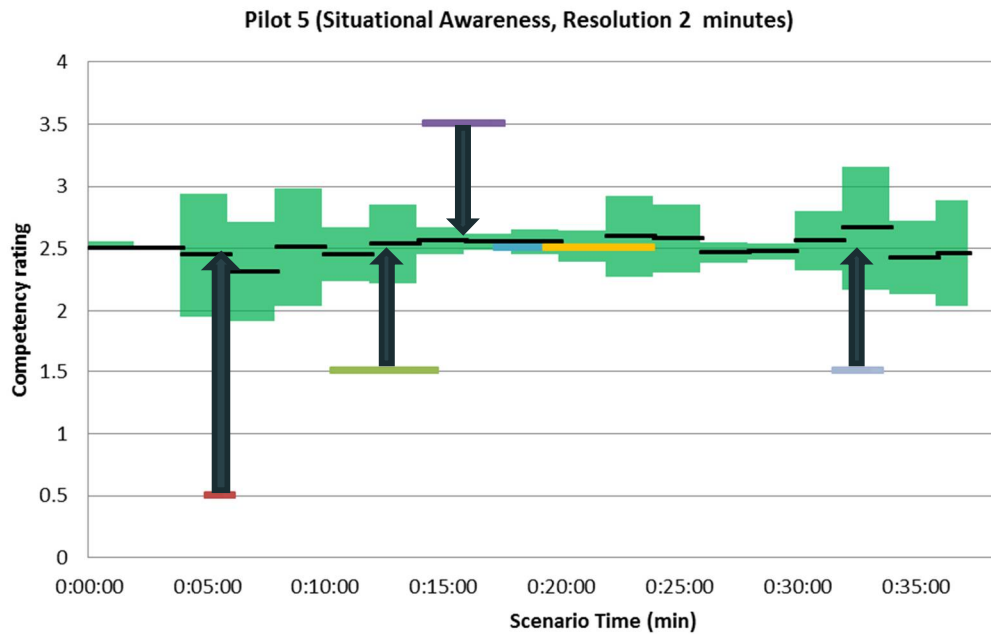
It is important that the distinction between desired and undesired mental representations (MR's) be based on some performance standard to prevent supporting the wrong (or less effective) set of mental representations. By coupling the competency assessment, specifically the TRA from section 4.2.2, with the mental representation of CATIE, the initial judgement of (un)desired MR's can be validated or adjusted. This is not to say that the competency assessment should be viewed as perfect, however it is a good source of validation as it is based on several expert assessments and also indicates where these assessments align well, or differ significantly in opinion.

Going back to Pilot 5 example, we can see how the MR analysis can be logically coupled with the competencies "Situational Awareness" and "Decision Making" as these are both related to the cognitive understanding of the situation. The competency "Application of Procedures" is a weaker link to the MR, and therefore not included in this validation exercise. Figure 79 below depicts the time based elements of the MR analysis, categorized as either Situational Awareness-related or Decision Making-related. Although the MR items seem sequential, they are in some cases in parallel, and in addition to this the time ranges are not exact. However, a good estimate is made for the time range in which that particular mental representation was either present or relevant.

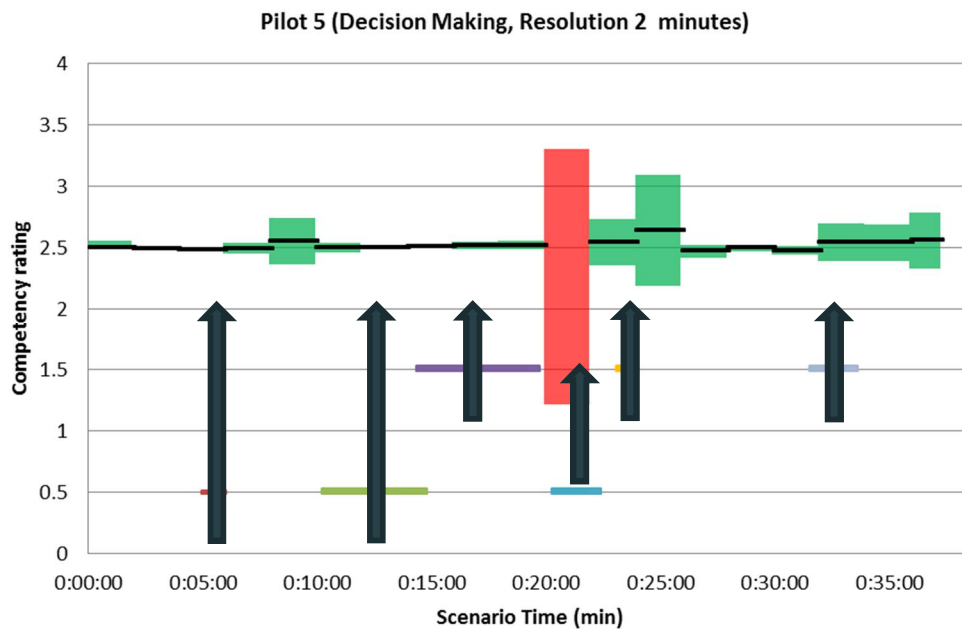
These time windows can then be plotted against the TRA data, as illustrated in Figure 80 and Figure 81. These plots show the TRA data in the same format as in section 4.2.2, as well as the MR items which are plotted under the TRA data, demarcated with diamond start-stop markers and colour coded in accordance with Figure 79. From these cross-plots, conclusions can be made concerning the proper appreciation of MR items, or if some MR's may require a different appreciation.

SITUATION AWARENESS	DECISION MAKING	
Not mention of fuel status	Does not identify there is a problem	Descent & Approach 27
Notice only fuel status and PF errors. No wind shift	Evaluates potential problems, identifies risk, considers alternatives	GO Around
Verbalize consequences for Landing distance. INOP problem and its consequences. He has an awareness of the aircraft state and its environment; he projects near future	Demonstrates knowledge of the failure but does not combine it with the operational status (new weather unknown)	AC BUS Failure
Accepts additional inputs from ATC. Mention CAT2 landing  Incomplete assessment of the situation: warning from LAPA	Does not decide on runway 09 as best option for landing  Passive in decision making, leaves decisions to ATC or PF: Declaring emergency  Misses elements: CWT limitations and its impact on landing	Second LAPA calculation
Require extra time to complete the picture of the situation  Has an awareness of the aircraft state in its environment.	Not searches in OMB. Misses elements: AP disengaged at 80ft  Accepts information and identifies the criticality of the situation. Select the best course of action when "ice on my window"	Approach

Figure 79: Overview of MERIA model items related to SA or DM



**Figure 80: Combining MERIA model SA-items (colour bars) with NLR's SA TRA**



**Figure 81: Combining MERIA model DM-items (colour bars) with NLR's DM TRA**

For the SA TRA + MR plot in Figure 80, four of the six MR items required an adjustment with regards to the consensus grading of the instructors. As the variance of this particular session is quite low, all ratings ended up on the same level (3 – acceptable).



For the DM TRA + MR plot in Figure 81, all MR items required an adjustment. Most adjustments were unambiguous, yet the fourth MR item (blue line, ~22 minutes) is adjusted only to level 2 – below acceptable. This is because it is the minimum level which can be warranted, despite the high level of uncertainty in the performance rating as indicated by the red band. The original MR level of 1 – unacceptable is too low, even when uncertainty is included.

This analysis has furthermore been completed for SA and DM competency performance measures for Pilots 3, 4 5 and 10 (these have multiple raters and TRA's). The MR items have been re-rated based on the average performance measures. Subsequently, in preparation for HMI design, the new MR items have been grouped per flight phase/event and their ratings summarized in the matrices below. From highest to lowest performance the colour representations are blue, green, light red, dark red. Grey areas are those where the competency performance was too uncertain to validate at all, and hashed areas indicate that the flight phase/event contained more than one MR items which had different ratings.

**Table 11: Overview of SA performance-corrected MR ratings per flight phase**

	SITUATIONAL AWARENESS PERFORMANCE					
Standard Phase	.....	Pilot 3	Pilot 4	Pilot 5	.....	Pilot 10
Descent & Approach 27						
GO Around						
AC BUS Failure						
Second LAPA calculation						
Approach						

**Table 12: Overview of DM performance-corrected MR ratings per flight phase**

DECISION MAKING PERFORMANCE						
Standard Phase	.....	Pilot 3	Pilot 4	Pilot 5	.....	Pilot 10
Descent & Approach 27						
GO Around						
AC BUS Failure						
Second LAPA calculation						
Approach						

These matrices summarise the validation of the mental representation items, and provide insight into the relative low-performance hotspots according to the MR nodes. The value of this analysis lies in the fact that HMI improvements should be focussed on the areas where low performance is an issue ("don't fix what isn't broken"). Four pilots is not nearly a statistically significant sample, yet these results do show that for many flight phases (in particular in the beginning) performance is high enough to remove focus on these flight phases. From an SA perspective a recommendation may be to improve SA support during the second LAPA calculation and Approach phase. From a DM perspective it may be valuable to design the new HMI with some form of decision support during the second LAPA calculation.

The matrix below shows the original MR item ratings, where the yellow boxes indicate that the rating has changed significantly due to the competency performance re-rating. It can be said that this validation has a significant effect on the original MR valuations, and should hold weight in the discussion which MR items may require support in the new HMI design.

**Table 13: Overview of DM performance-corrected MR ratings per flight phase**

	.....	Pilot 3	Pilot 4	Pilot 5	.....	Pilot 10
Descent & Approach 27						
GO Around						
AC BUS Failure						
Second LAPA calculation						
Approach						

#### 4.2.4. Considerations for HMI design in support of Pilot Monitoring

In general, from cross-analysis of Scenario 2 we have observed several difficulties for the pilots:

- To recognise changes in the internal or external situation (new weather, availability of airports and tracks etc.). Missing communication with the ATC may induce to lose some possible solutions in critical situations.
- To understand the impact of these changes on the landing procedure (wind shift meaning change of runway, inoperative systems changing the type of landing and the application of new procedures, etc.).
- To anticipate problems or future scenarios, for example to know the amount of remaining fuel at landing, to know the time consumed by procedures, etc.
- To recall and monitoring all landing parameters as the weather changes, runway distance, failure type, inoperative systems and in the same time to perform procedures that relate all these parameters.

The analysis of the 3 main aspects (fuel status, electrical failure, weather) impacting the mental representation of the critical situation provides the following conclusions:

- Fuel status is critical and impacts the time in flight. The actual representation of fuel status is in kilograms, but it could be enriched with new considerations such as distance, remaining time, remaining "legal" time and so on. By improving pilot's awareness of fuel status it is possible to help him/her to:
  - Have a better understanding of the possible options;
  - Improve the prioritization of actions, and consequently have a better time management;
  - Have a better communication with the captain and ATC to take decisions.

- Pilots had difficulties in understanding the importance of different elements in the ECAM status and their relations with the environmental conditions while managing the Electrical failure. In particular, it is advisable to integrate consistent communication of weather information (in particular in relation to available/possible runways and airports) and consequences of the failure in terms of aircraft limitations. To perform this consistent communication, 3 representations needs to be included in the ECAM:
  - simple representation of situation;
  - simple representation of possibilities;
  - simple representation of consequences of choices.
- Weather. HMI should help pilots to have a better representation of the weather, and to be sure that this representation sticks to reality. The actual code to communicate the weather conditions is not spontaneously understood in terms of impact. As important as instant weather condition, weather tendency is needed to anticipate on near future, especially during critical phases.

## 5 CONCLUSIONS AND RECOMMENDATIONS

### 5.1. Conclusions

This report presents the results of several analyses and correlation tasks, thus many conclusions can be drawn based on the outcomes of each task and each section.

Going back to the structure of the deliverable reported in Section 1.5 (see Figure 2) the three main goals of this document are:

- To prove the HPE model in a partially controlled simulation setting, using data collected during Scenario 1. The proof of concept is divided into two parts:
  - One dedicated to the correlations between runs and HPE factors (measured through subjective ratings), runs and physiological factors, and runs and performance (Section 2);
  - The other dedicated to finding connections between performance and HPE, between HPE and physiological factors and finally to link the two relations in order to analyse the correlation between performance and physiological factors and identify a potential equation to predict the performance (Sections from 3.1 to 3.3).
- To test the HPE model in an ecologically valid setting, basically taking the abovementioned predictive equation based on Scenario 1 and trying to apply it to Scenario 2, with a different task and different performance measures (Section 3.4).
- Finally, to use the results from the previous sections to identify performance decrement areas and improve HMI to support pilot's performance recovery.

From the first set of analyses, it emerged that the runs designed for Scenario 1 were actually able to affect the intended factors – for example, Run 3 designed to produce high workload in the subjects was able to do that. However the ecological experimentation in flight conditions did not allow the gradual increase of factors as it can be done in a full laboratory setting. Thus the difference between high workload condition and very high workload condition was not experienced as such.

The experiment indicates links between three of the components which are supposed to shape the human envelope: the increase of workload is associated with an increase of the stress level and a decrease of the situation awareness, and these changes happened in all the runs of Scenario 1. As the three parameters all interrelate, the modification of the experimental conditions to change the level of one parameter has always side effects on the two others. However, it seems that the runs designed to push the workload affected the HPE more than the others, getting pilots closer to the edge of their performance. In other words, an increase of workload had a stronger impact on the stress level and situation awareness than the other way around.

Despite the reduction of the envelope comes always with an increase of the normalised heart rate and of the normalised mean eye radius, the correlations between each individual factor (stress, workload and situation awareness) and these physiological markers cannot be precisely evaluated with these experiments as all the factors are always affected in each run. The other physiological parameters

measured in these simulations, the heart rate variability (measured with SDNN), showed inconsistent trend, as it seems to be modified not in the same way by the workload increase than by stress or degraded situation awareness. The analysis of physiological data finally confirmed that it is difficult to find physiological markers for degraded situation awareness alone, as both the stress and the workload levels were also modified by the experimental condition. Also, often the operator is not aware that his/her situation awareness is degraded, and this reflects in no variation in his/her physiological markers.

Looking at the correlation between runs and performance, it seems that the envelope was more constrained in the workload conditions than in the conditions affecting stress and situation awareness, with higher deviations from glideslope and localiser and higher number of go-arounds, indicating that being closer to the edge of the envelope degraded pilots' performance, pushing them to interrupt the approach in order to recover higher safety margins. However, it has to be noted that results on performance has to be taken carefully, as the conditions of some runs (i.e. level of turbulence and localiser interferences) had a direct impact on the performance parameters.

Another outcome of the first set of correlation tasks is that the combinations of factors show a marked effect on performance (objective and subjective), higher than the single factor effect even at Medium-Medium-Medium levels. Runs 7 and 8 demonstrated that the combined effect of the three HPE factors reduced more severely the envelope, with performance more critically affected as shown by the very frequent use of go-around manoeuvre to recover safety margins - even in the cases in which the manoeuvre could be seen as a questionable decision (e.g. in the low fuel condition). A summary of each factor in each run is reported in Table 14. Finally, results on physiological factors in these two runs did not defend the hypothesis that the combination of workload, stress and situation awareness factors played a different role on pilot's psychophysiological values.

The analysis of the eye tracking data and hotspots proved to be a valid method to give insights on pilot's situation awareness, as it is able to indicate what he/she was doing and looking and how the flight crew reacted at key points. This enables the analyst to make certain inferences about the information that is important, and what is comprehended and carried forward. As the other physiological measures, it is sensitive to the dynamics of the situation, so the granularity of analysis is important.

**Table 14: Summary of correlation tasks between runs and HPE / Physiological data / Performance data**

Single Factor Effect				
High / Very High Workload condition (Run 3 – Run 4)	HPE	Workload ↑↑↑	Stress ↑↑	Situation Awareness ↓↓
	Physio	Heart Rate ↑↑↑	Heart Rate Variability ↓	Pupil diameter ↑↑
	Performance	LOC Deviation ↑↑	G/S Deviation ↑	% Go Around ↑↑
High Stress condition (Run 5)	HPE	Workload ↑	Stress ↑↑	Situation Awareness ↓↓
	Physio	Heart Rate ↑↑↑	Heart Rate Variability ↑↑↑	Pupil diameter ↑
	Performance	LOC Deviation ↑	G/S Deviation ↑	
Highly degraded SA condition (Run 6)	HPE	Workload ↑↑	Stress ↑	Situation Awareness ↓↓
	Physio	Heart Rate ↑↑	Heart Rate Variability ↑↑	Pupil diameter ↑↑
	Performance	LOC Deviation ↑↑↑	G/S Deviation ↑↑	% Go Around ↑
Combined Factors Effect				
Medium WL / Medium Stress / Medium SA condition (Run 7)	HPE	Workload ↑↑↑↑	Stress ↑↑↑	Situation Awareness ↓↓↓
	Physio	Heart Rate ↑	Heart Rate Variability ↓	Pupil diameter ↑↑↑
	Performance	LOC Deviation ↑↑↑	G/S Deviation ↑↑↑	% Go Around ↑↑↑
High WL / High Stress / High SA condition (Run 8)	HPE	Workload ↑↑↑↑	Stress ↑↑↑	Situation Awareness ↓↓↓
	Physio	Heart Rate ↑↑	Heart Rate Variability ↑↑	Pupil diameter ↑↑↑↑
	Performance	LOC Deviation ↑↑↑↑	G/S Deviation ↑↑↑	% Go Around ↑↑

On the other side, when we tried to move from these results to a “predictive” approach the results became blurred. The second set of correlation tasks tried to build a relation between Performance – HPE – Physiological data able to use the physiological data as predictors of performance decrement. The generated equation was then validated through its application to a different scenario. To end up with this predictive equation, we studied the correlation between Performance and HPE and then the correlation between HPE and physiological factors. Despite significant correlations emerged in both analyses, these only explained small part of the variation, meaning that the relations exist but they are weak, or the equations don’t take into account other relevant factors that play a bigger role in the HPE variation.

Despite these weak results, we generated a predictive equation that expressed the performance as a function of the physiological factors associated to each element of the HPE [ $Performance = 0.445 \times HR + 0.278 \times EYE + (0.130 \times HR + 0.139 \times EYE) \times (0.130 \times HR + 0.133 \times EYE) \times (-0.126 \times HR)$ ]. This equation was applied to Scenario 2, using the physiological data from that scenario and the performance calculated with the competency performance ratings. The results from this analysis indicate a very weak relation at best between predicted and competency performance, indicating that the variations in the competency performance are not sufficiently explained or mirrored by a change in the predicted performance.

In the end, the mathematical construct for proving the HPE could not successfully translate from Scenario 1 to Scenario 2 and cannot be used to identify performance decrements or develop adaptive interfaces based on physiological monitoring. However, there are multiple possible explanations for this invalidation. First and foremost, the performance metric in Scenario 1 (flight path deviation) and the performance metric in Scenario 2 (competency measures) are quite different in what they observe, and as such do not necessarily correlate. Secondly, the prediction formula is a mathematical construct designed around the multi-factor HPE concept, but also permits accumulation of errors. As such the model is possibly prone to sensitivity, which is somewhat visible in the predicted performance dataset, which has several major peaks and valleys. Third, the scoped HPE concept using only three factors (workload, stress & SA) may not cover all the facets of performance, and therefore be limited in its predictive power. Lastly, the analyses that generate the equation from Scenario 1 rely on the global levels of workload, stress and situation awareness from the top of descent to the decision altitude. The analysis cannot reflect or predict sudden or short changes in the level of these parameters. Moreover, even if the physiological data have been normalised, their relationship with HPE factors are certainly partly subject dependant. The study of the links between changes of HPE factors for small duration (within a scenario) and changes in physiological data has not been addressed here. It would require a more continuous evaluation of stress level and situation awareness.

Despite the validation tasks didn’t give the expected results, the self-assessed performance and experts’ analysis allowed the identification of the areas of intervention for HMI improvements. With respect to Scenario 1, changes should be put in practice for the Electronic Flight Bag, Navigation Display and Primary Flight display, which should integrate wind information from the ground. The implementation of Head Up Display in the cockpit could facilitate the collection of relevant aircraft parameters such as speed, altitude, glide slope, flap settings and wind. For sure, pilots would be helped by on-board visualisation of ground-



related information during landing phase, as well as by a system able to support performance calculation by correlating aircraft parameters/configuration with environmental information. The latter in particular is expected to facilitate the remaining fuel calculation, a critical task for pilots. Some suggestions on preferred warning channels were also collected in the debriefing phase. Here, pilots show a strong preference towards visual channel for non-critical communications or “kind warnings” (far before the situation becomes dangerous), while the audio channel should be limited to the critical warnings that require an immediate intervention.

With respect to Scenario 2 results, several difficulties for pilots were observed. Among others, problems in recognising changes in the internal or external situation (new weather, availability of airports and tracks etc.) in a timely fashion, in anticipating problems or future scenarios, and in recalling and monitoring all landing parameters as the weather changes, runway distance, failure type, inoperative systems and in the same time to perform procedures that relate all these parameters. From the cognitive walkthrough three main aspects impacted the mental representation of the critical situations: fuel status, electrical failure and weather. Some recommendations were thus derived:

- Fuel status is critical and impacts the time in flight. The actual representation of fuel status is in kilograms, but it could be enriched with new considerations such as distance, remaining time, remaining “legal” time and so on.
- Pilots had difficulties in understanding the importance of different elements in the ECAM status and their relations with the environmental conditions while managing the Electrical failure. ECAM could be improved by creating a simple representation of the situation, of the possibilities and of the consequences of the different choices.
- HMI should help pilots to have a better representation of the weather, and to be sure that this representation sticks to reality. The actual code to communicate the weather conditions is not spontaneously understood in terms of impact. As important as instant weather condition, weather tendency is needed to anticipate on near future, especially during critical phases.

## 5.2. Recommendations

In the next round of simulations, to be held in Bordeaux in the autumn of 2017, the key lessons learnt from this report have to be taken into account:

1. It is possible to create linear models to predict subjective ratings of Workload and Stress from a variety of physiological measures, mostly heart rate and pupil diameter.
2. However, not all physiological measures are equally valuable at the prediction.
3. Eye tracking can give us useful information on situation awareness degradation and how the interface is used.
4. To understand the HPE in general, other models should be used, as the ones from this report were able to explain only a modest amount of variation.

5. The equation linking physiological measures to performance is task dependent, thus cannot be used in a predictive way

In the end, as the HPE independent factors can be expressed as a function of the physiological signals, and physiological measures are able to predict some variation, the use of these signals cannot be completely discharged for the next simulations. However, attention has to be paid on how the data are used (normalised per pilot, for factors validation instead of performance prediction) and the conclusions that can be derived from them, especially on subject's situation awareness.

HMI improvement has to tackle issues that may affect pilot's situation awareness, such as fuel status and weather conditions (in particular, wind shift and/or sudden weather changes). The electrical failure seems to be a more contingent issue, related to the specific scenario under investigation, so ECAM improvement can be tackled separately from the previous two issues.

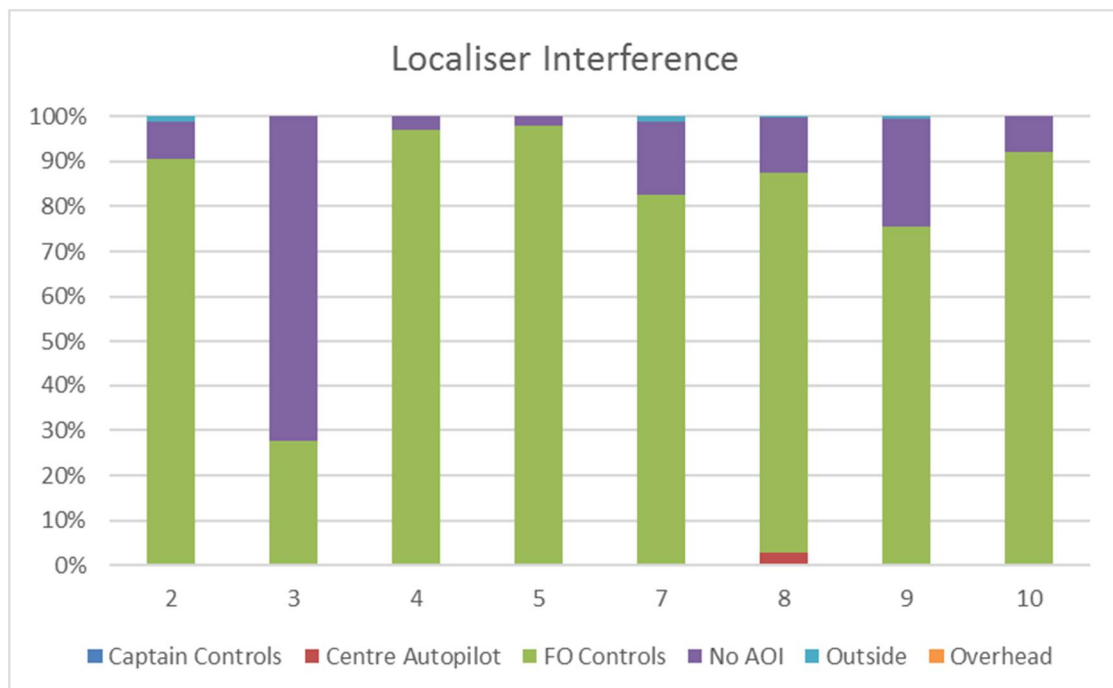
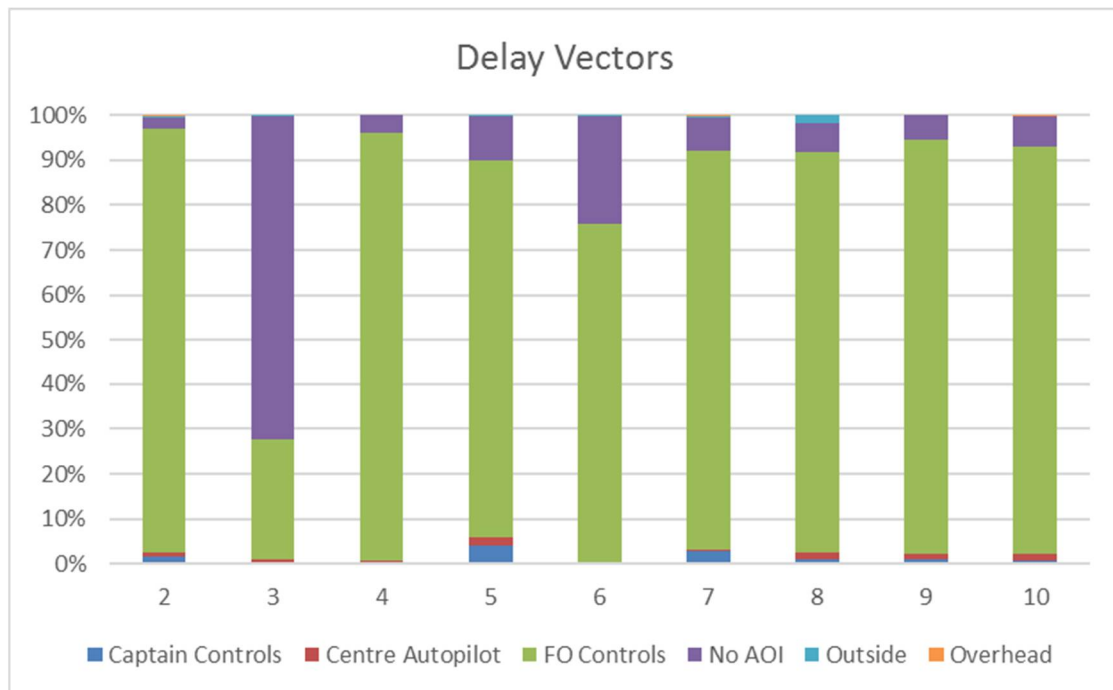
This document may be updated on basis of the outcomes of the second round simulation experiments as planned in Future Sky Safety P6. This update, with the latest findings, is planned at the end of the project.

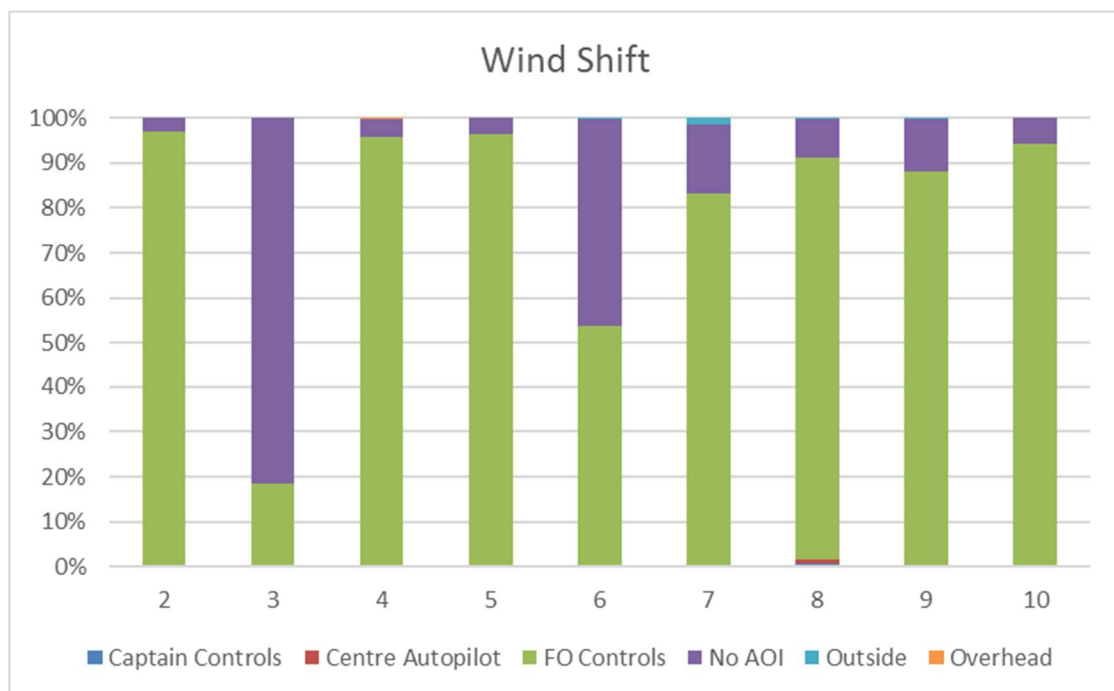
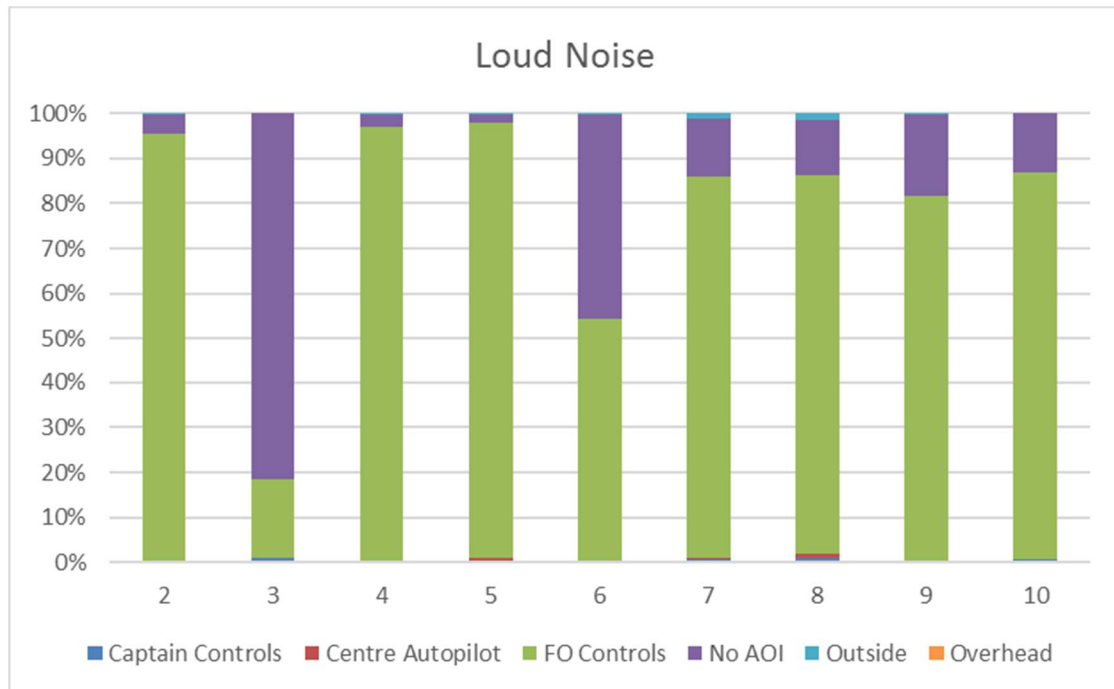
## 6 REFERENCES

- Callan, D. J. (2016). Eye Movement Relationships to Excessive Performance Error in Aviation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.
- Duchowski, A. (2007). *Eye Tracking Methodology: Theory and Practice*. Springer Science & Business Media.
- Ellis, K. (2009). Eye tracking metrics for workload estimation in flight deck operations. *Theses and Dissertations*. Retrieved from <http://ir.uiowa.edu/etd/288>
- 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), 32–64.
- Glenstrup, A., & Engell-Nielsen, T. (1995). *Eye controlled media: Present and future state*.
- Jacob, R., & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*.
- Nakayama, M., & Shimizu, Y. (2004). Frequency analysis of task evoked pupillary response and eye-movement. *Proceedings of the Eye Tracking Research & Applications Symposium on Eye Tracking Research & Applications - ETRA'2004*, 71–76. <http://doi.org/10.1145/968363.968381>
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '00* (pp. 71–78). New York, New York, USA: ACM Press. <http://doi.org/10.1145/355017.355028>
- Stephane, A. L. (2012). Eye tracking from a human factors perspective. In G. A. Boy (Ed.), *The Handbook of Human-Machine Interaction: A Human-Centered Design Approach*. Ashgate Publishing, Ltd.
- Wickens, C., Xu, X., Helleberg, J., & Marsh, R. (2001). Pilot visual workload and task management in freeflight- A model of visual scanning. *Focusing Attention on Aviation Safety*. Retrieved from
- Yu, C.-S., Wang, E. M.-Y., Li, W.-C., Braithwaite, G., & Greaves, M. (2016). Pilots' Visual Scan Patterns and Attention Distribution During the Pursuit of a Dynamic Target. *Aerospace Medicine and Human Performance*, 87(1), 40–7. <http://doi.org/10.3357/AMHP.4209.2016>

## Appendix A SITUATION AWARENESS AND EYE TRACKING DATA

### Appendix A.1 Stacked Bar Charts for Key Events





## Appendix A.2 Flight deck dialogue Pilot 6

Start time	End time	Communication / event
0:10	0:11	Auto-pilot off
0:30	0:31	Clarification of airport
0:49	1:10	Capt: "shall I look for something on the page, if any specialties?" FO: "please" Capt: "Frankfurt VR, VR1. Approach 25 left and Charlie second set" Capt: "OK like this?" FO: "looks good thank you"
1:37	1:40	Capt: "I ask for the latest weather report, OK?"
1:40	2:14	Capt comms with ATC
2:16	2:21	FO: "Flaps?" Capt: "Flaps full"
2:30	2:55	Capt comms with ATC. Confirmation of any gusts
2:56	3:36	FO: "Max X wind components is within the limits?" Capt: "Its 38 isn't it?" FO: "yes" Capt: "I check both winds. If they are parallel they are reliable, otherwise not, and if they are reliable I take a look at the cross wind and then you make some inputs to the rudder"
3:37	3:53	Descent from 110 to 70. Confirmation between Capt and FO
4:32	4:53	FO: "flight directors off?" Capt: "yep"
5:23	5:30	FO: "Switch it on" Capt: "2000 a little bit high but as long as he isn't saying anything"
5:56	6:03	FO: "are we ok for the transition" Capt: "Yes" FO: "ok" Capt: "I verify"
6:03	6:27	Capt gets verification from Frankfurt approach. Capt sets new vectors
7:40	8:01	Capt asks FO about heading and whether to set a waypoint
8:20	8:38	Instruction from ATC to reduce speed. Continue on D616. Capt enters info into MCDU.
8:43	8:55	Capt changes entry using EFIS. ATC alerts them that they are number 15 to land
8:55	8:59	FO: "fuel on board remaining 1500kgs"
8:59	9:14	FO: "I'm not sure we can make it that long. What do you think? With the fuel flow 700 kilos that's 20 minutes. I don't think we can be number 15 with the remaining fuel"
9:15	9:29	FO: "700 kilos an hour? 2400" Capt: "just tell me, what do you think?"

Start time	End time	Communication / event
9:30	9:35	FO: "maybe you can ask the controller for priority due to remaining fuel?"
9:36	10:00	Capt: "we have a fuel flow of 2400 / hr if we fly straight and level, so we have 36 minutes remaining. Is that correct?" FO: "yes" Capt: "so I ask for priority?" FO: "Can you ask him for an estimated approach time for us?"
10:01	10:27	Capt contacts ATC, confirms 25 mins flight time.
10:27	10:55	FO: "we will have to use some of our reserve fuel so then we have to declare mayday" Capt confirms that even if they turn now they have less than 1200 Capt: "so it's an emergency now isn't it?"
10:56	10:58	FO: "so declare mayday"
10:58	11:13	Capt declares mayday due to low fuel. Present time remaining 35 minutes. Requests priority for the approach.
11:14	11:29	Capt inputs new vectors
11:32	11:36	FO: "can you switch on the landing lights please?"
11:38	12:02	Capt contacts ATC to confirm low fuel. Told they may speed up
12:02	12:08	Capt confirms heading and speed
12:10	12:34	FO: "it doesn't make sense to speed up. Maybe we can do 250 for a while?" Capt sets speed top 250. ATC gives descent information. Capt confirms
12:56	12:58	Capt: "2000 *FOs name*?" FO: "yep"
13:01	13:04	Capt: "speed is 280 almost"
13:11	13:16	FO: "speed indicator is very erratic" Capt confirms
13:20	13:30	ATC gives new vectors. Capt confirms
13:31	13:46	Capt inputs direct into MCDU. FO: "that's a normal direct. Thank you" Capt: "Localiser info is too far out" FO: "OK"
13:52	13:53	Capt: "1000"
14:06	14:16	Capt: "wind up here is between 22 and 40 knots. Not very much cross" FO: "That's good"
14:34	14:53	Capt: "Did we read the altimeters?" Capt confirms readings FO: "speed reducing"
15:22	15:27	Capt: "I have no localiser information. You too" FO: "localisers working"
15:28	15:38	Comms from ATC to follow localiser. Cleared ILS 25L



Start time	End time	Communication / event
15:40	15:55	Capt: "approaching 1200 kilos of fuel" FO: "that's half an hour" Capt: "shall I turn on the flight directors again? Maybe it helps you?" FO: "I would like to fly without the directors"
16:01	16:55	Capt: "ceiling was 400 feet. Do we have to think on anything for the go around?" FO: "in case of go around, it will be more time critical than even now because we are low on fuel, and if we do the go around it's to the south so we are not really terrain critical" Capt: "watch your altitude" FO: "thank you" Capt: "Perhaps I ask for short radar vectors?" FO: "that would be a good idea" Capt: "We have 2 minutes to letkey, so that means 1100 kilos, so 26 minutes remaining. I ask for very short vectors"
16:58	17:09	Capt has comms with ATC. Change of tower.
17:12	17:35	FO: "manage speed" Capt confirms vectors and speed
17:36	17:52	Capt: "In case of go around, if we started from the runway we will have 24 minutes, so we need very short radar vectors back"
17:54	18:07	FO: "speed ok for the moment. Flaps 1" Capt: "speed checked" FO: "flaps 1"
18:10	18:14	FO: "select speed 140" Capt sets speed
18:25	18:38	Capt: "Correct glideslope approx 9 miles FFM. I don't have an ILS"
18:30	18:30	<b>EVENT: TOD glideslope</b>
18:38	18:41	FO: "flaps 2" Capt: "speed is checked"
18:53	19:00	Capt: "would you like any autobrake?" FO: "autobrake medium is a good idea"
19:02	19:06	Comms from ATC informing number 1 for landing 25L
19:09	19:12	Capt: "wind 250, 32 knots"
19:18	19:22	FO: "what was the wind again?" Capt: "250, 32 knots"
19:27	19:41	FO: "set go around at altitude 5000ft" Capt: "I check both winds, they are completely parallel"
19:41	21:20	<b>EVENT: Loud noise</b>
19:41	20:22	Capt: "what is this?" FO: "I don't know. Can you hear me?" Capt: "I read you. I cannot turn it off" FO: "I don't know what it is. But the engines are working, the aircraft is flying" Capt: "landing clearance is missing, and landing configuration"
20:24	21:51	<b>EVENT: Wind shift</b>

Start time	End time	Communication / event
20:25	20:26	Clear to land
20:26	20:28	Gear down
20:46	20:51	<i>Capt: "Flaps 3. Speed checked"</i>
20:56	20:59	<i>Capt: "flaps full. Speed checked"</i>
21:02	21:07	<i>FO: "select approach speed"</i>
21:17	21:21	<i>FO: "landing checklist"</i>
21:24	21:25	<i>Capt: "landing, no blue"</i>
21:27	21:40	Capt comms with ATC. Confirm still clear to land.
21:44	21:56	<i>Capt: "just wanted to have brake level here confirmed to land. The 1000 is checked"</i> <i>FO: "ground contact indicator still parallel?"</i> <i>Capt: "yes"</i>
21:57	22:06	Capt confirms windspeed. <i>Capt: "max crosswind no problem"</i> <i>FO: "yes"</i>
22:16	22:22	<i>Capt: "clear to land. Landing checklist completed. Everything is done"</i>
22:25	22:33	<i>Capt: "watch your glideslope"</i>
22:40	22:43	<i>FO: "we seem to be offset to the runway"</i>
22:44	22:48	<i>Capt: "I see the approach lights"</i> <i>FO: "correcting"</i>
23:07	23:08	<i>FO: "runway in sight"</i>
23:11	23:12	<i>Capt: "power slightly high"</i>
23:16	23:16	<b>EVENT: Decision altitude</b>
23:27	23:30	<i>Capt: "Centreline! Centreline! Centreline! Brake!"</i>
23:31	23:31	EVENT: Touch down
23:45	23:46	Autobrake off.

## Appendix A.3 Detailed deep dive and heat-map analysis

Until 2:16 the Captain is asking questions. The FO is required to agree, which requires a passing glance, an acknowledgement which would indicate level 1 SA. He is not required to make any decisions, only agree with them.



**Figure 82: AOI frequency and direction pilot 6 from 0 to 2:16 minutes**

As Figure 82 shows, the majority of the FOs gaze was toward the PFD (AOI 17), fluctuating between the PFD and the ND (AOI 16), in a constant monitoring state. An interesting pattern occurred after the captain asked if he should check his MCDU. In this instance, the FO moved his gaze from his ND, to the ED, to the SD, to the captains RMP, to his own RMP. This scan pattern would indicate that the FO is checking the vectors and cross checking confirmation prior to agreement with the captain developing L2 SA from the perception of elements in the L1 scan.

From 2:16 to 2:21 the first officer asked the captain the state of the flaps. This did not require the FO to move their gaze from the PFD, even after confirmation of “flaps full” from the captain. This would indicate that still at this stage in the flight, the FO only has to perceive information. In a sense, the FO is offloading his SA requirement to the Captain who is required to perceive and comprehend the information in order for the FO to agree.

From 2:56 to 3:36 The FO is again asking the captain questions regarding the weather. The FO spends the majority of his time during this segment looking at his PFD (see Figure 83) but also glanced first at the SD, then the ED. However, although these glances indicate that they perceive information, and maybe comprehend the future state of the aircraft.



**Figure 83: AOI frequency and direction pilot 6 from 2:56 to 3:36 minutes**

During this time, the captain demonstrated that he was reaching level 3 SA by using the information (wind direction and speed), checking it was a certain way (parallel, based on experience) then using this to guide future decisions (rudder). The FO was checking the information based on the captains decisions.

From 3:37 to 3:53 the FO and captain confirm a level change, from FL110 to FL70. This is again led by the captain, and the FO uses this information to make the relevant changes to the flight. During this time, the FO keeps his focus confined to his controls (figure 16) again indicating that they are acting on the information that he has just received.



**Figure 84: AOI frequency and direction pilot 6 from 3:37 to 3:53 minutes**



From 4:32 to 6:03 there is another exchange between the captain and FO, with the FO leading the questioning. Again, from the dialogue it appears that the FO is asking questions about the current state of the aircraft, rather than the future state, requiring confirmation from the captain. It is the captain who is required to perceive and comprehend the information in order to pass the status on to the FO. However, the FO's gaze deviates from his two main panels during this time (Figure 85).



**Figure 85: AOI frequency and direction pilot 6 from 4:32 to 6:03 minutes**

During this 90 second section, the FO glances at the ED and SD as well as the MCDU and Radio Management Panels (RMP). This may indicate a 'double check' or a glance at the panels that he is asking the captain to relay the information from. The FO is reaching level 2 SA at this point, but so far has not shown any indication of using this information to project the future state; this has been carried out by the captain.

The two minutes between 6:03 and 8:01 sees a reversal in information flow between the captain and FO. The captain, who has been communicating with Frankfurt ATC is thinking about the endpoint; the landing. The captain asks the FO if they should set a waypoint, to which the FO says yes. Interestingly, despite this being a two minute segment, and the FO being required to make a decision about their future heading, the FO remains focussed on his ND and PFD (Figure 86).



**Figure 86: AOI frequency and direction pilot 6 from 6:03 to 8:01 minutes**

8:43 into the scenario, ATC alerts the flight crew that they are number 15 to land. At 8:55 the FO first notices that there is low fuel (Figure 87).



**Figure 87: AOI frequency and direction pilot 6 from 8:55 to 8:59 minutes**

The FO is focussed on the ED for a large proportion of this 4 second segment. Following this initial realisation, the FO, for the first time during this flight demonstrates that he is reaching level 3 SA. From 8:59 to 9:35 the FO uses the information gained from the ED to establish the future state regarding fuel levels. This can be seen in Figure 88.



**Figure 88: AOI frequency and direction pilot 6 from 8:59 to 9:35 minutes**

The FO concludes that the remaining fuel left will not allow them to be number 15 to land. The FO still looks to the captain to confirm that the calculation is correct, and then suggests asking for priority. This continues to the next segment, but the captain assigns the decision making back to the FO, who decides the best approach would be to ask for an estimated approach time. During this time, the majority of the FO's focus is on the ED (see Figure 89).



**Figure 89: AOI frequency and direction pilot 6 from 9:36 to 10 minutes**

The FO is using the information to predict the future state of the flight. This is a role reversal from earlier in the flight where the captain was the one making the crucial decisions. The fact that the FO spends the majority of his focus on the ED rather than his PFD would also indicate that the FO's focus has switched



from solely flying the plane to making decisions. This is especially apparent when compared to figure 18, when the FO was making a decision about waypoints; his focus remained on the ND and PFD, as it was a fairly arbitrary decision. The decision regarding the fuel however required the FO to make calculations using current information and project to the future, indicating level 3 SA.

Despite the low fuel situation, the FO does not declare that it is yet an emergency situation. The FO has comprehended the information, but his projection (that they are ok for now, but they may have to declare an emergency) is different to the projection of the captain (that they should declare emergency). This misalignment of views could be due to a number of things, including the FO being embarrassed that he did not notice the fuel situation until it became an emergency, or the FO calculating that the 36 minutes remaining fuel would be sufficient for the 25 minutes approach time estimated by ATC. In this case, it is possible that the FO is behaving reactively, whereas the captain is looking at the situation proactively and considering all of the eventualities; if they continue and are then required to do a go around, for instance. In addition, the FO is still flying the plane manually so much of his effort is focussed on this. Figure 90 shows that the FO's focus is still on the ED, as well as the SD and the captains ED.



**Figure 90: AOI frequency and direction pilot 6 from 10:27 to 10:58 minutes**

The FO agrees with the captain eventually and asks him to declare an emergency. Following this, the captain inputs new vectors and receives confirmation from ATC that they may speed up. The FO declares that it does not make sense for them to speed up, but then agrees to do so for a while. This segment is shown in Figure 91.



**Figure 91: AOI frequency and direction pilot 6 from 12:10 to 12:34 minutes**

The FO is seen to focus on the centre autopilot, as well as the main controls. While the FO is comprehending the information, the captain is still making the decisions, with the agreement of the FO. This pattern of the captain asking and the FO confirming continues until 14:34, with the FO focus remaining mainly on the ND and PFD. At 14:34, information provided by the captain requires the FO to take action. This segment is displayed in Figure 92.



**Figure 92: AOI frequency and direction pilot 6 from 14:34 to 14:53 minutes**

The captain provides the FO with speed information, to which the FO reacts. This can be seen by the scan pattern in figure 24. At 15:40 the captain gives an update on the fuel situation, declaring 1200 kilos of fuel left. The FO is seen to confirm this himself by glancing at the ED (Figure 93), indicating a high level of SA.



**Figure 93: AOI frequency and direction pilot 6 from 15:40 to 15:55 minutes**

The FO uses this information to calculate that they have half an hour remaining. During the following segment, (16:01 to 16:55) the captain once again takes the lead in initiating decision making with the FO. The FO confirms that if a go around is required, it is not terrain critical. This conclusion / decision was based on existing route knowledge rather than the perception and comprehension of current information, but this guides the captain to make a decision regarding asking ATC for very short radar vectors.

The captain continues to initiate decision making until 19:09 when the captain confirms the wind direction and speed. At this point, the FO is focussed on his ND and PFD (Figure 94)



**Figure 94: AOI frequency and direction pilot 6 from 19:02 to 19:12 minutes**



Although the wind information is available to him, and he is looking at the relevant displays, the FO asks the captain for re-confirmation of the wind information. This would indicate that the FO is concentrating on the task of flying the aircraft, which is confirmed in the next segment when his attention is solely on his PFD.

At 19:41, the scripted loud noise starts. The captain acts surprised and glances around the cockpit. The FO however reacts calmly and initially checks that the captain can hear him. He then checks his own instrumentation (Figure 95).



**Figure 95: AOI frequency and direction pilot 6 from 19:41 to 20:22 minutes**

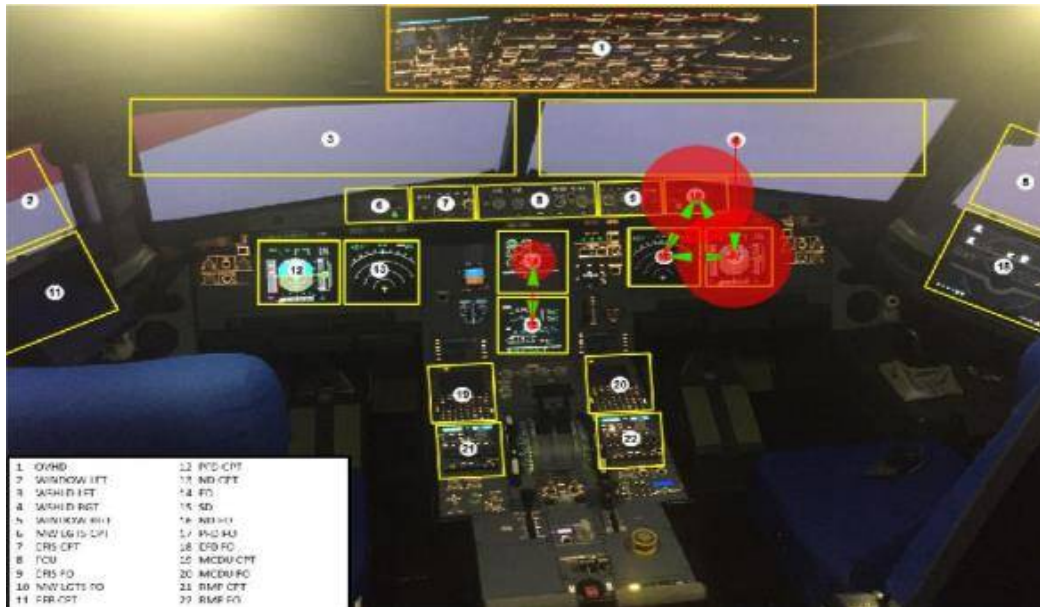
During this loud noise event, the FO, as Figure 96 shows, checks his own instruments to ensure that there is nothing drastically wrong with the aircraft. By monitoring his ND, PFD, ED, SD, MCDU and RMP he is able to comprehend the current situation and confirm it through the instrument readings. He confirms to the captain that the 'engines are working and the aircraft is flying in order to carry on with the task at hand. There is no indication either through the dialogue or the eye tracking data that the FO is using this information to do any more than confirm the current situation, or reaching level 2 SA.



**Figure 96: AOI frequency and direction pilot 6 from 20:25 to 21:25 minutes**

Following the loud noise the crew continue to prepare for the imminent landing and proceed with gear down and the landing checklist. This is initiated by the FO and requires confirmation by the captain. The FO's scan pattern (Figure 97) indicates that he is monitoring the flight instruments in order to administer the landing checklist and maintain consistent flight, suggesting a good level of SA.

During the final approach, the captain confirms that the landing checklist is complete and everything is done. The FO then confirms that they are slightly offset from the runway centreline. During this segment the FO's attention is focussed solely on the PFD which is to be expected given the challenging manual control of the aircraft. This sustained gaze toward the fused information about the aircraft status on the PFD would maintain SA across the three levels in the Endsley model.



**Figure 97: AOI frequency and direction pilot 6 from 23:07 to 23:31 minutes**

During the final landing segment shown in Figure 97 (23:07 to 23:31), the majority of the FO's gaze is divided between the PFD and the MW lights, and there appears to be a scanning loop between these two instruments and the ND. This is expected during the final landing phase, and the captain feeds information to the FO throughout this segment.