	**HUMAN** Model-based Analysis of Human Errors during Aircraft Cockpit System Design	

D4.8 - Requirements for the cognitive model after cycle #2

Project Number: **Nature:** **Classification:** **Version:** **Parts & Classifications:** **Work Package(s):**	211988 Deliverable Public Main document (Public) WP4
Document Timescale: **Start of Document:** **Final version due to:** **Time now:**	Project Start Date: March 1, 2008 T0+34 T0+36 T0+38
Issue Date (dd/mm/yyyy):	29/04/2011
Compiled:	Tina Mioch (TNO)
Authors:	Rosemarijn Looije (TNO), Tina Mioch (TNO)
	Jan-Patrick Osterloh (OFF), Florian Frische (OFF)
Technical Approval:	Tina Mioch (TNO)
Issue Authorisation:	Andreas Lüdtke (OFF)

DISTRIBUTION LIST		
Copy type[1]	Company and Location	Recipient
T	HUMAN consortium	All HUMAN Partners

[1] Copy types:
M = Master copy,
E = Email,
C = Controlled copy (paper),
D = Electronic copy on disk,
T = TeamSite (Sharepoint)

RECORD OF REVISION			
Version	Date (dd/mm/yyyy)	Status Description	Author
V0.1	19-10-2010	Initial structure	RL
V0.2	10-01-2011	Added content to Analysis H32	RL
V0.3	17-02-2011	Added content to Analysis H32	RL
V0.4	05-04-2011	Added content to Analysis H17/H36	RL
V0.3	22-04-2011	Added content to Introduction, methodology etc.	TiM
V1.0	29-04-2011	Last review	TiM

1 Table of Contents

2 Abbreviations and Definitions

Abbreviation	Definition
LC	Learned Carelessness
WM	Working Memory
CL	Cognitive Lockup
CASCaS	Cognitive Architecture for safety Critical Task Simulation
PSP	Physical Simulation Platform
VSP	Virtual Simulation Platform

3 Introduction

The main objective of this document is to describe the requirements for the cognitive model derived from the results of the experiments of the second cycle. More specifically, it is described how the data of the second cycle has been analyzed, and the requirements that we derive from these results.
In the following section, the methodology of deriving the requirements is described, followed by the description of the analysis of the simulator data and the requirements derived from this analysis. Thereafter, the error production mechanisms for the second cycle are specified and requirements are derived from this specification. At the end, an overview of the different requirements is given.

4 Methodology to determine the requirements

The following methodology is used to determine the requirements for the cognitive model from the second cycle:

- Analysis of subjective data, which are the questionnaires filled in by the pilots flying and monitoring
- Analysis of the EPMs
  - On the PSP
  - On the VSP (if relevant)
  - Comparison of the results of the analyses on the PSP and VSP (if relevant)
- Analysis of requirements regarding the EPMs and associated ETs for the future.

5 Selection of Hypotheses to be Evaluated

In D4.3, we described in detail the method of how we derived at the possible hypotheses. For the list of hypotheses, see Table 1. Here, you also find which of the hypotheses were chosen to be evaluated in cycle 1, and which were chosen in cycle 2. For a description of the method that we used to decide which hypotheses to evaluate, please see D4.3. For the second cycle, we used the same method to decide on the relevance and importance of the hypotheses. We also took the experiences from the first cycle into account. For example, for hypothesis 9 (*There is a common scanning pattern which depends on the flight level*), no significant difference could be found in the first cycle for the scanning patterns in different flight levels, and no new requirement was specified for the cognitive model. We thus assume that there is no different scanning pattern, and will not further investigate this hypothesis in the second cycle.

ID	Hypotheses	1st cycle	2nd cycle
1a	Every individual pilot has a standard order (which is different from other pilots) for checking the displays (Uncover pattern in pilot traces, e.g. via data mining techniques)	No	No
1b	Every individual pilot has a standard order (which is different from other pilots) for checking the displays (Define patterns beforehand and verify/refuse these patterns based on search in pilot traces)	Yes	No
2	Every individual pilot's scanning pattern depends on the workload	No	No
3	Every individual pilot's scanning pattern depends on the flight phase	No	No
4	Individual pilots have individual transition points (for flight level) in relation to the scanning pattern	No	No
5	Every individual pilot's scanning pattern depends on the flight level	No	No
6	Pilots have common orders (general scanning patterns) for checking the displays (Define patterns beforehand and verify/refuse these patterns based on search in pilot traces)	Yes	Yes
7	Common scanning patterns depend on the workload	No	No
8	Common scanning patterns depend on the flight phase	Yes	Yes
9	Common scanning patterns depend on the flight level	Yes	No
10	Distribution of gaze on AOIs is not significantly different for all pilots	Yes	Yes
11	Distribution of gaze depends on flight level	Yes	No
12	Distribution of gaze depends on flight phase	Yes	Yes
13	Distribution of gaze depends on workload	No	No
14	Reaction time of pilots to visual events (AHMI popup box) does not differ significantly	Yes	Yes
15	Reaction time of pilots to visual events (AHMI popup box) depends on flight level	Yes	No
16	Reaction time of pilots to visual events (AHMI popup box) depends on flight phase	Yes	No
17	Reaction time of pilots to visual events (AHMI popup box) depends on workload	No	Yes
18	Individual pilots have individual transition points (for flight level) in relation to the reaction time	No	No
19	Task completion time (AHMI related tasks) does not differ significantly for pilots	Yes	Yes
20	Task completion time (AHMI related tasks) increases if pilots perform other tasks in parallel	No	No
21	The more often a pilot performs a certain task (AHMI related tasks), the less time he will need to complete the task	Yes	Yes

22	Pilots perform tasks (AHMI related tasks) by applying "our" normative procedures (comparing the actions of all pilots to the formalized normative procedures)	No	No
23	Certain events trigger the same procedures for all pilots (two aspects: does a particular trigger lead to the same procedure for all pilots? And: do all pilots show the same actions for a particular procedure (not necessarily the normative behaviour as we modelled it))	No	No
24	If a task is suspended or interrupted by pilots, the other task has at that moment a higher priority	No	No
25	The AHMI related tasks are of a higher priority than the monitoring task	No	No
26	Multi-Tasking behaviour depends on attributes, such as flight level, flight phase, work load	No	No
27	If pilots perform AHMI related tasks they will suspend the monitoring task	Yes	No
28	If procedures are alternative, they have the same p-value (for all task alternatives, did the pilots choose them alternatively, i.e. did some of them have a different order of execution than others)	No	No
29	Individual pilots will have individual p-values for alternative procedures (for an individual pilot, does it hold that this pilot chooses always the same alternative above another)	No	No
30	Pilots have the same goals in certain situations (e.g. after having received an event)	No	No
31	Every action a pilot executes belongs to an active goal	No	No
32	At moments in the scenario where high workload is expected, the pilots subjectively experience a high workload.	No	Yes
33	If the CL is busy with a goal, the sign-symbol translation may be delayed (e.g. too late for an appropriate response).	No	No
34	If two signs resemble each other on some of their dimensions, the risk of confusing the signs is higher.	No	No
35	Pilots will show Learned Carelessness in the "Handle Uplink" procedure after 15 repetitions (Pilots will not perform the check of the Constraint Flight Level in vertical mode)	No	Yes
36	The pilots will show Cognitive Lockup at moments in the scenario when there are multiple tasks with similar priorities when the pilot is executing a task with a high mental workload	No	Yes (hypothesis is changed)
37	If the CL is busy with a goal, and the AL needs help with the execution/planning of another goal, the CL reacts later to this call than if it was not already busy.	No	No
38	The CL should monitor the AL at certain points, but gets a high priority task which it focuses all attention on, not monitoring the AL at the control points.	No	No
39	The individual differences have interactions with each other	No	No

	(someone that changes behaviour around FL… also checks the displays in a specific order)		

Table 1: List of hypotheses relevant for HUMAN (cycle 1 and cycle 2)

6 Analysis of Hypotheses about the Basic Capabilities

In this section we will present the results of the analysis of the *basic capabilities* of CASCaS for the selected hypotheses. First analyses on basic capabilities have been conducted during cycle 1. The results have been used to validate the performance of CASCaS and great effort has been made to improve the performance of CASCaS. During the second cycle, the analyses have been repeated in order to verify the results from the first cycle and finally to demonstrate the improvements of CASCaS with a new set of reference data. Hypotheses related to the basic capabilities are the following:

- **H6:** Pilots have common orders (general scanning patterns) for checking the displays (Define patterns beforehand and verify/refuse these patterns based on search in pilot traces)
- **H8:** Common scanning patterns depend on the flight phase
- **H10:** Distribution of gaze on AOIs is not significantly different for all pilots
- **H12:** Distribution of gaze depends on flight phase
- **H14:** Reaction time of pilots to visual events (AHMI popup box) does not differ significantly
- **H19:** Task completion time (AHMI related tasks) does not differ significantly for pilots
- **H21:** The more often a pilot performs a certain task (AHMI related tasks), the less time he will need to complete the task

6.1 Analysis of H6

In order to analyse if a common scanning order for all pilots exists we have analysed the scanpaths taken by the pilots in the cockpit. We defined a scanpath as the transition from one AOI in the cockpit to another AOI. In the first cycle we have investigated 2-series (A → B) and 3-series (A → B → C) scanpaths. Because pilot interviews revealed that monitoring scanpaths do not depend on past scanning actions only 2-series scanpaths are analysed during cycle 2. Analyses have focused on a set of 6 AOIs (PFD, HSI, AHMI, EFCU, ENG, and Window). Thus, our analysis considers 30 (6^2-6) scanpaths (self-transitions excluded). For each experimental run a distribution profile has been derived from the eye tracker data. The profile represents for each scanpath how often a scanpath has been selected in relation to the sum of all scanpaths in percent. The profiles of each pilot during the scenarios have been aggregated and then compared to the average profile of the whole population (exclusive the pilot analysed) in order to measure the correlation

(Pearson r) of each individual pilot to the population. In Table 2 the results of the correlation analysis are presented.

Pilot	Cruise	Approach	Final Approach
1	0.97	0.98	0.9
2	0.9	0.89	0.98
3	0.98	0.91	0.96
4	0.86	0.90	0.88
5	0.86	0.97	0.96
6	0.99	0.97	0.96
7	0.78	0.86	0.9
8	0.98	0.99	0.97
9	0.96	0.98	0.98
10	0.89	0.94	0.97
11	0.96	0.89	0.98
12	0.96	0.99	0.97
13	0.8	0.89	0.76
14	0.97	0.99	0.96
15	0.98	0.95	0.96
16	0.9	0.98	0.97

Table 2: Correlation values (Pearson r) between each pilot and the population per flight phase

It can be seen that the individual performance of each pilot highly correlates with the group performance. The results demonstrate that a commonality of order for checking displays is highly probable. The aggregate scanpath profile of all pilots for the different flight phases is depicted in Figure 1.

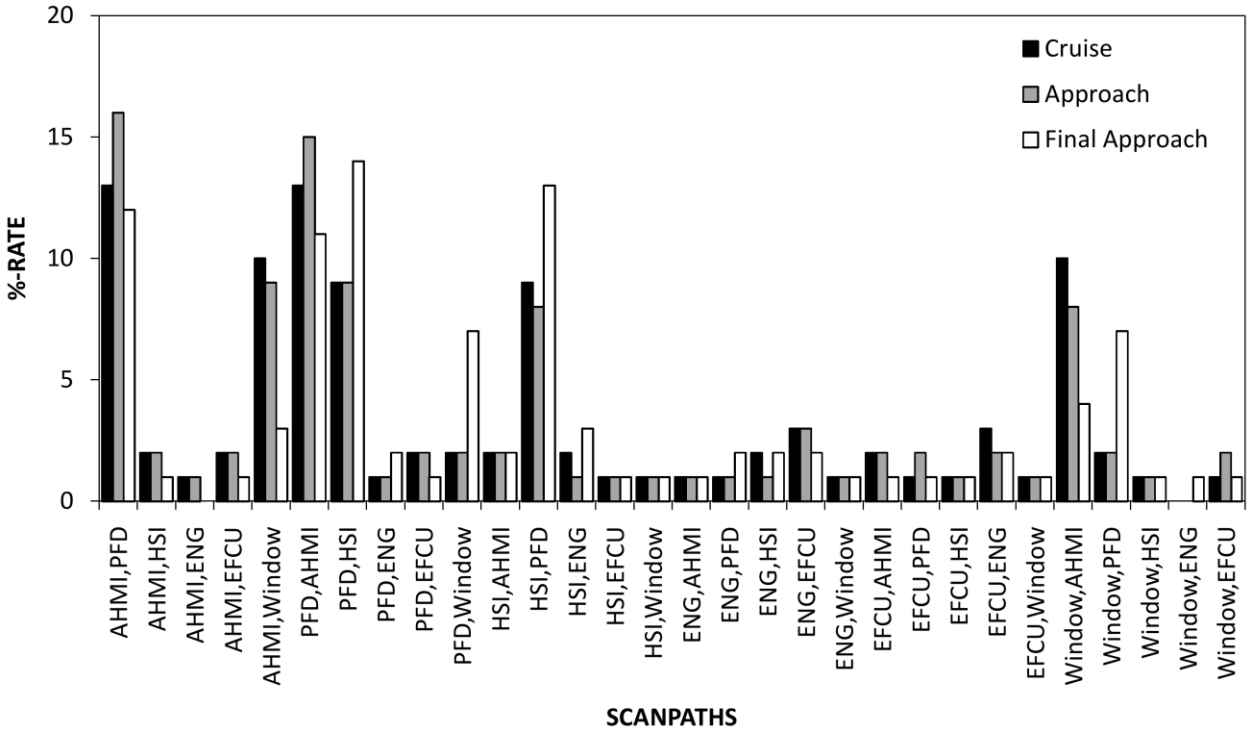

Figure 1: Scanpath profile of human pilots during cruise, approach and final approach

Interviews with pilots revealed that no normative scanning procedure (like the "Basic T" scan) exists which pilots apply for their scanning activities under normal flight conditions. Thus, pilots decide on their own in which order displays are scanned. However, the scanpath profiles depicted in Figure 1. reveal that differences exist between the flight phases. These differences are related to the flight tasks which differ during the phases. While during cruise phase, organizational tasks are dominant (such as organizing the flight route, monitoring weather and traffic) during approach and final approach flight tasks (monitoring flight parameters relevant for landing) are dominant. Thus, the way of how pilots scan in the cockpit depends on the flight tasks rather than on a directive. Differences between scanning behaviour in "normal" operation during cruise and during the AHMI Uplink task have been analysed in order to test, if the AHMI influences the scanning behaviour of pilots. For both cases, the two most dominant scanpaths (primary and secondary path) have been used to draw a transition graph for the AOIs investigated. The results of normal operation during cruise are depicted in Figure 2 and the results of the AHMI task analysis are depicted in Figure 3.

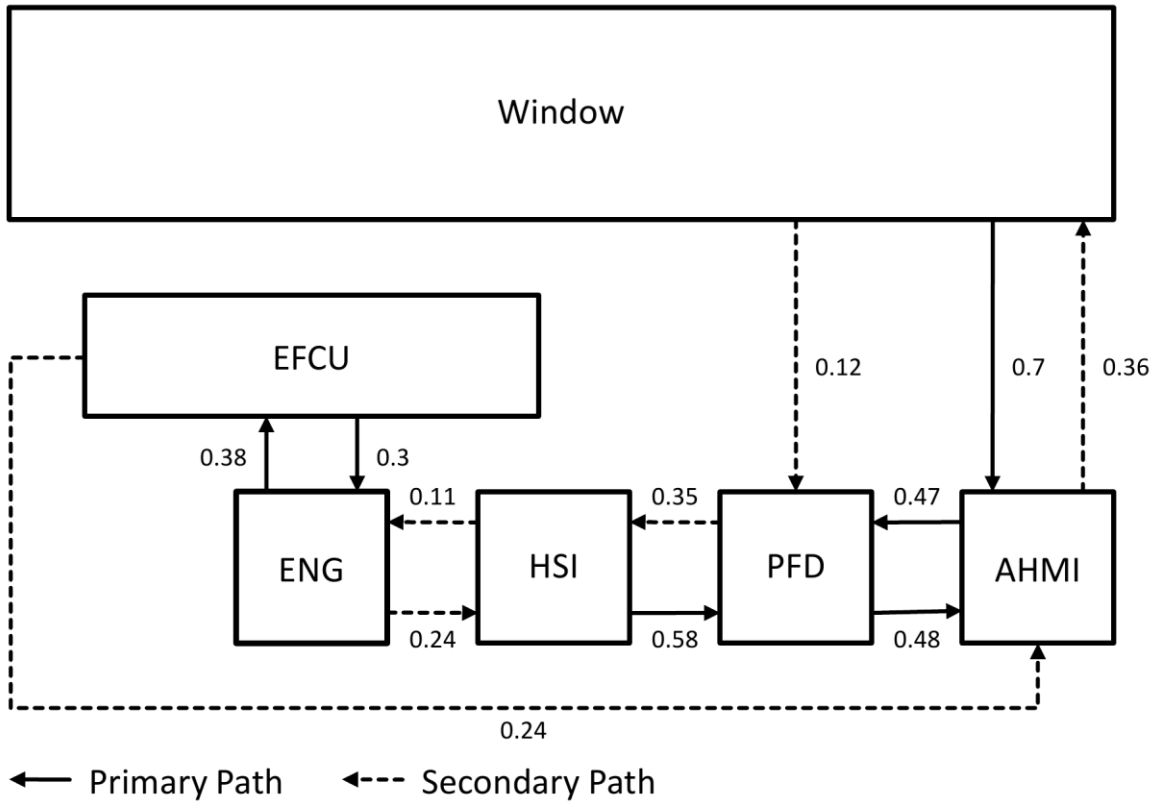

Figure 2: Scanpaths (primary and secondary) of human pilots during monitoring task while flying in cruise phase (numbers on transitions are transition probabilities)

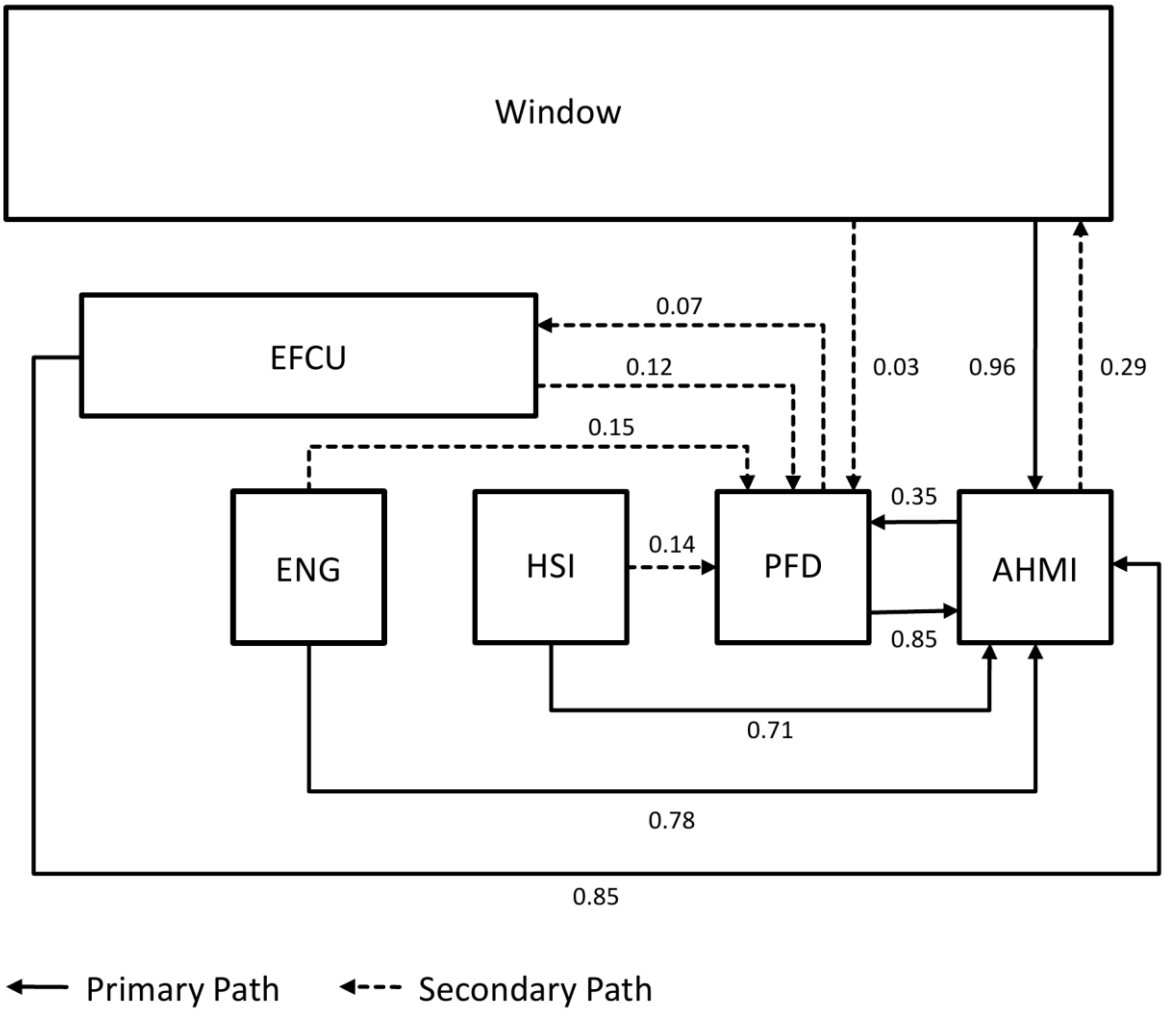

Figure 3: Scanpaths (primary and secondary) of human pilots during AHMI uplink task while flying in cruise phase (numbers on transitions are transition probabilities)

The numbers on each transition between two AOIs represent the probability for a transition from one AOI to another AOI. In this analysis, scanning is considered as a markow process of first order, meaning that past scanning behavior does not affect decisions for future transitions. The results for the scanpaths analysis during monitoring confirm the results of cycle 1 analysis, which is that pilots optimize their scanning behavior with regard to the cockpit layout (see also D4.3). The analysis of scanpath during AHMI task reveals that pilots tend to perform transitions back to the AHMI. Thus, the attention focus during the task is mainly centered on the AHMI which is what we already stated in D4.3.

We have compared the results of human scanning behavior with results of model behavior. For the comparison of the scanning behavior between model and human

pilots during monitoring task in cruise, approach and final approach, see Figure 4, Figure 5 and Figure 6.
Results reveal a large consistency between model and human pilots for the behavior during different flight ($r \geq 0.9$). In addition, the scanpaths of human pilots and the model during the AHMI uplink task have been compared. Results are depicted in Figure 7 and Figure 8. The performance analysis shows that behavior during the AHMI task is not that consistent and tuning might be useful ($r \leq 0.5$).

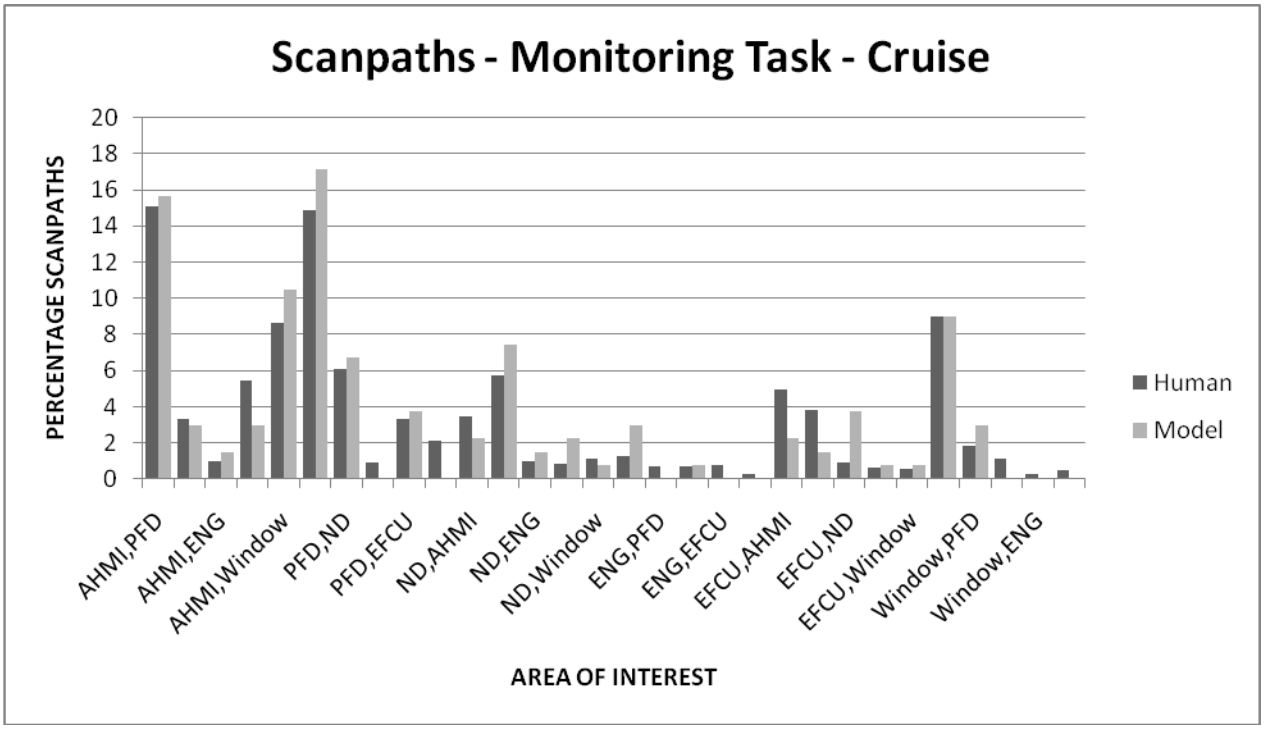

Figure 4: Comparison of scanpaths between model and human pilots during monitoring task in cruise

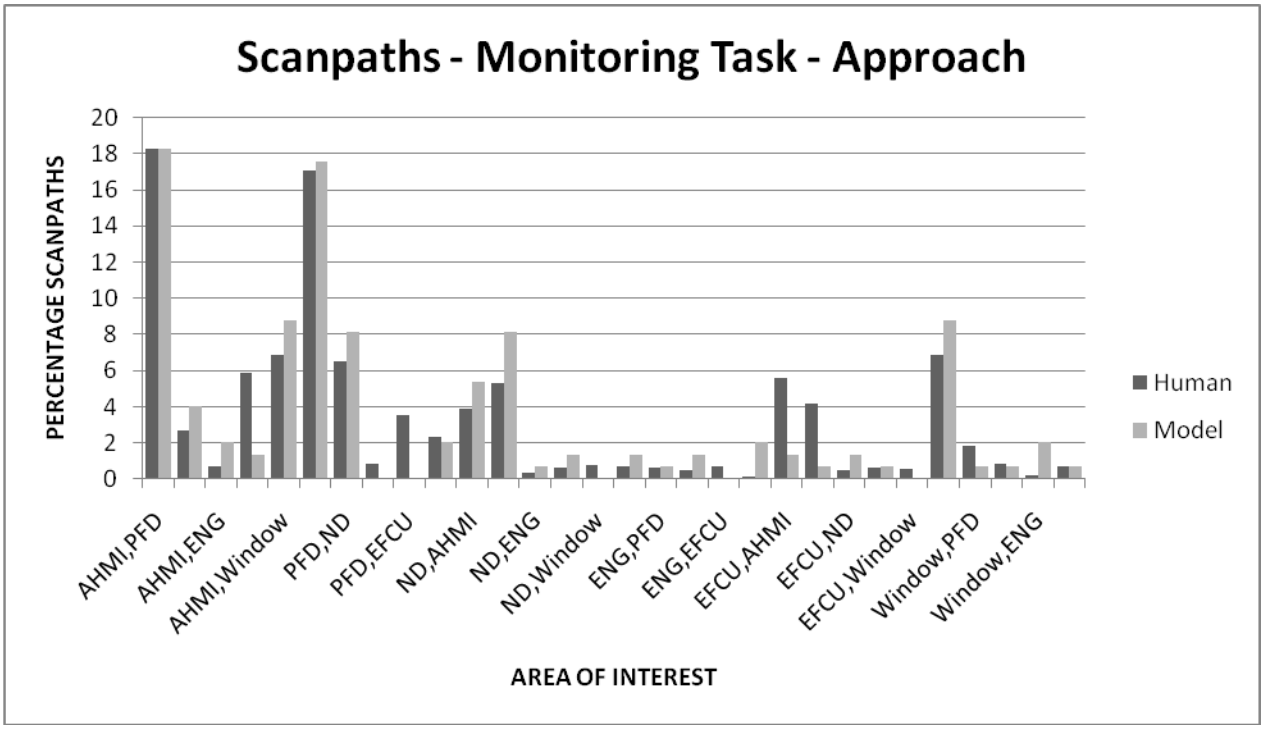

Figure 5: Comparison of scanpaths between model and human pilots during monitoring task in approach

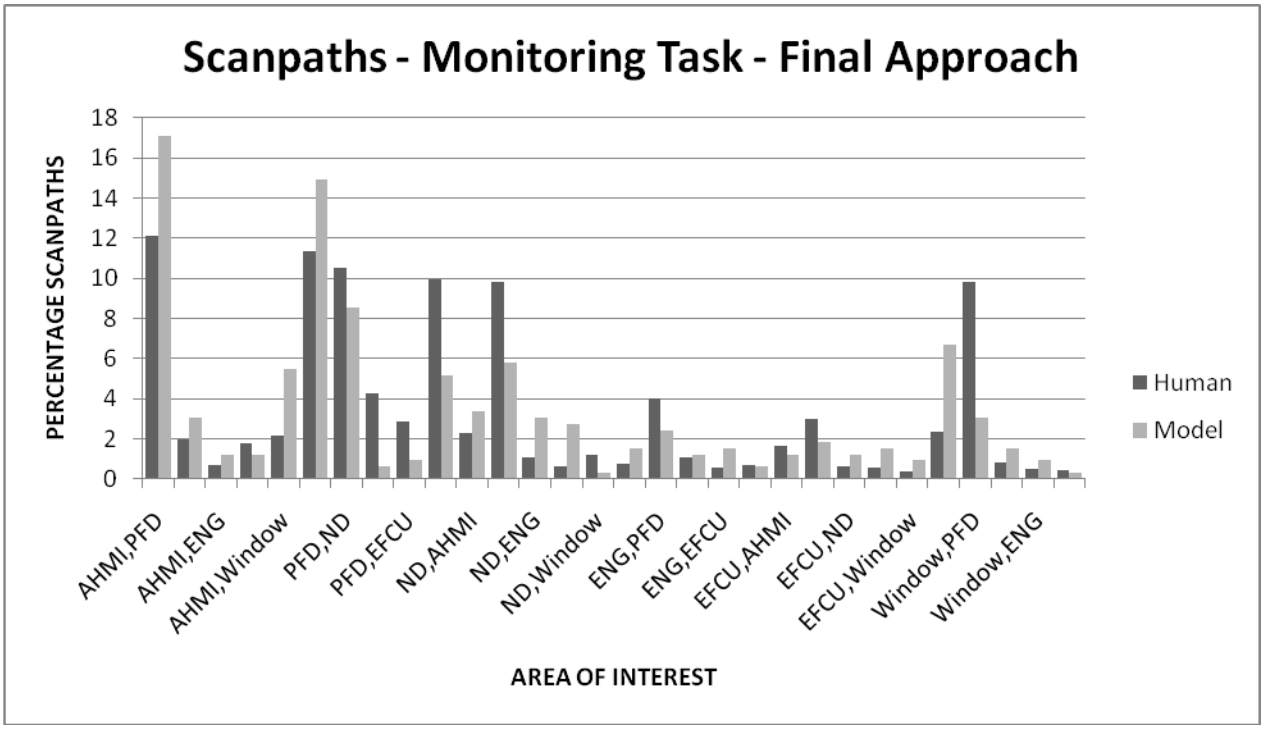

Figure 6: Comparison of scanpaths between model and human pilots during monitoring task in final approach

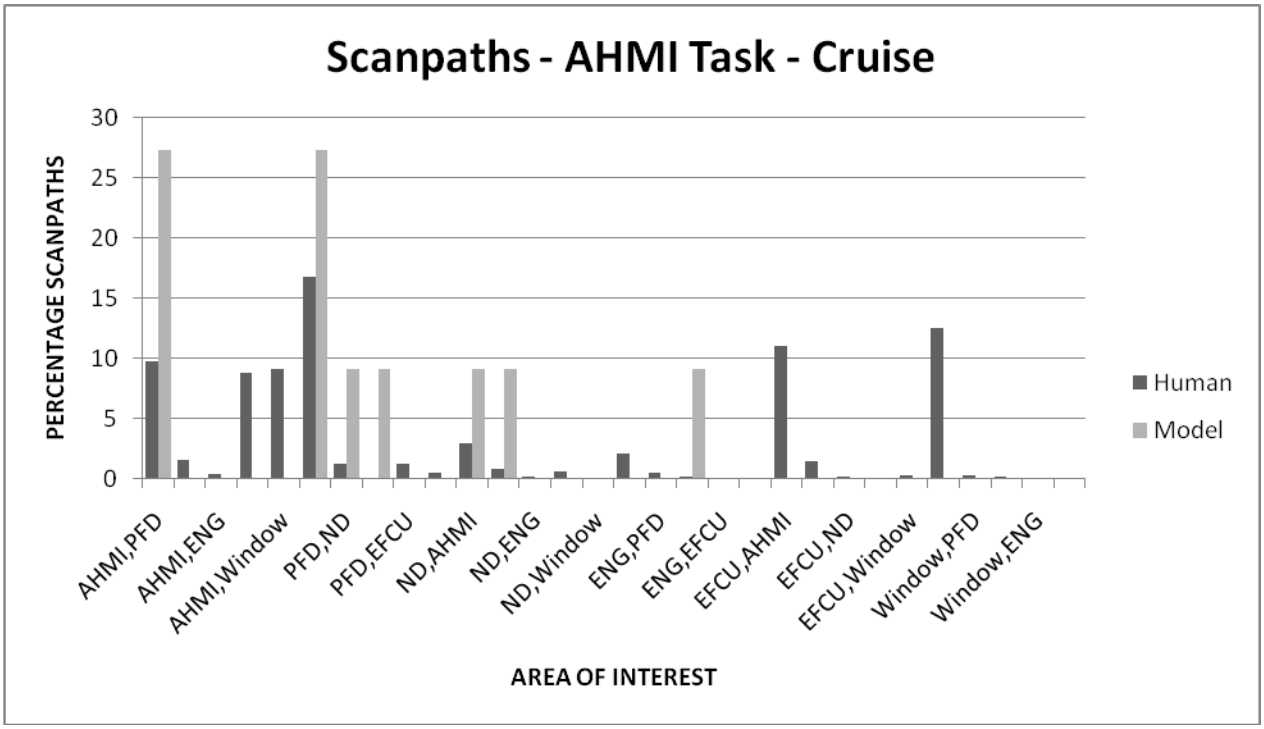

Figure 7: Comparison of scanpaths between model and human pilots during AHMI uplink task in cruise

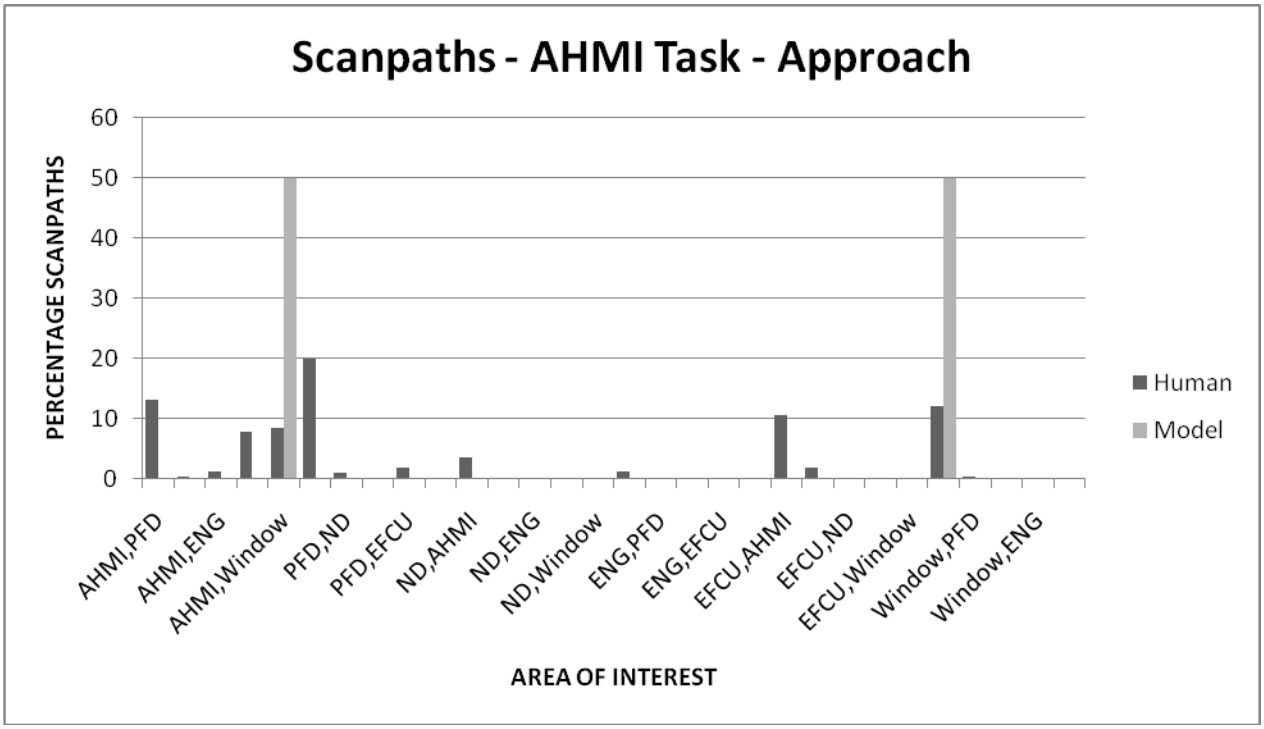

Figure 8: Comparison of scanpaths between model and human pilots during AHMI uplink task in approach

6.2 Analysis of H8

The scanpath profiles of human pilots during cruise, approach and final approach have been compared and a correlation analysis (Pearson r) has been performed. Results are depicted in Table 3.

	Cruise	Approach	Final Approach
Cruise	1	0.98	0.76
Approach	0.98	1	0.79
Final Approach	0.76	0.79	1

Table 3: Correlation values (Pearson r) between the flight phases

Results show that correlation is lowest for comparisons where final approach phase is involved. E.g. the correlation between cruise and final approach measured with r=0.76. However, a significance test revealed that *overall* differences between all phases are not statistically significant ($p>0.05$) although there are remarkable differences, e.g. with regard to transitions between AHMI and window. The reason for differences in partial scanpaths can be seen in task switching and task priorities during the flight phases. Pilots reported about organizational tasks during cruise

phase, e.g. tasks related to the AHMI, monitoring of weather and traffic. On the other hand, pilots reported about flight tasks, especially during final approach, where monitoring the aircraft state is much more relevant. AOIs related to these differences are mainly AHMI, PFD and the window. A scanpath analysis on this subset of AOIs reveals a significant change ($p<0.05$) during cruise and final approach, and also between approach and final approach.

6.3 Analysis of H10

[i]Analyses of the first cycle have on gaze distribution have shown that the overall gaze distribution follows a general trend (rank order of AOIs during flight phases). In the second cycle the analyses have been repeated in order to verify the results from the first cycle. Figure 9 depicts the gaze distribution results of the second cycle, which have been recorded during the experiments conducted in the GECO.

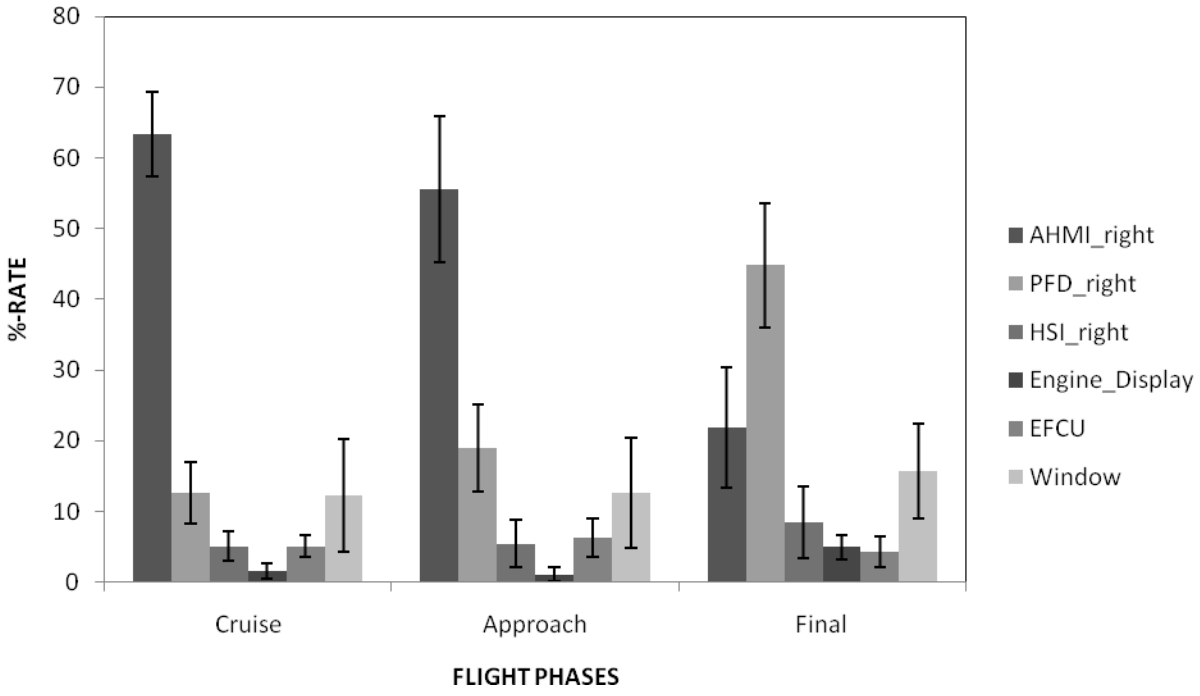

Figure 9: Gaze Distribution of Human Pilots during Flight Phases

The error bars reveal that individual differences exist for the exact performance values. A significance test has been performed to compare the individual performance of each pilot with the aggregate performance of all pilots in order to asses if the degree of differences. The test shows that the individual differences are statistically significant ($p<0.05$) for all flight phases. However, an analysis of rank

orders as performed during cycle 1 confirms the results of cycle 1. Monitoring activities depend on the pilot's overall understanding of the flight tasks during the flight phases. The rank order analysis shows that, in general, pilots have a common understanding of these tasks but the concrete execution plan on the level of actions varies between pilots.

A comparison between model and human performance reveals that the model's variability has improved during the first and second cycle. However, in order to cover performance of different pilots more variability is needed. A direct comparison between model and human mean performance is depicted in Figure 10. It can be seen that mean performance between model and human pilots is very well ($r \geq 0.9$) for all flight phases. However, the variability in model performance varies in average 5% around each measuring point, which is too small, in comparison to the human performance results.

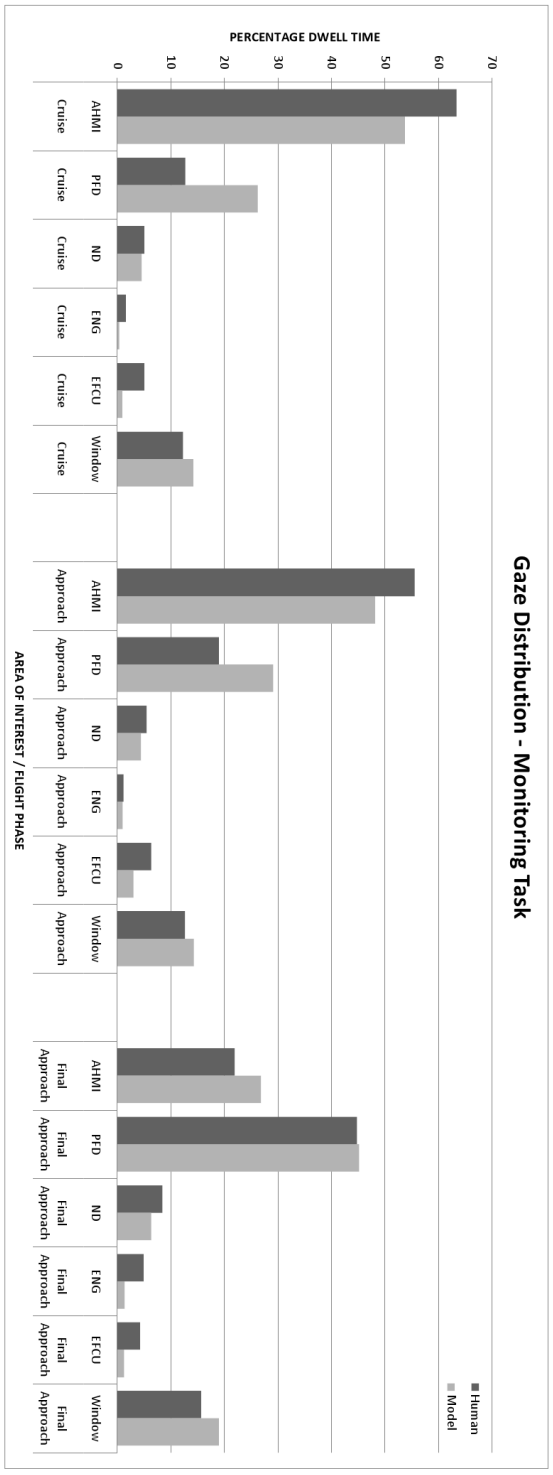

Figure 10: Gaze Distribution during flight phases, comparison between model and human pilots

6.4 Analysis of H12

Significant differences exist between cruise and final approach, and approach and final approach ($p<0.05$). Minor differences exist between cruise and approach. Distribution of gaze is depicted in Figure 9. This result is in line with the analyses performed during cycle 1.

6.5 Analysis of H14

Human pilot's reaction time to the AHMI task is around 1 second in average and is mainly dependent on the distance between the eye fixation location and the AHMI message box location at the moment of message box popup. Larger distances can result in not perceiving the message box ad hoc, but in later stages when the message box comes into the visual field of the pilots. In Figure 11 the average reaction time per subject pilot is presented.

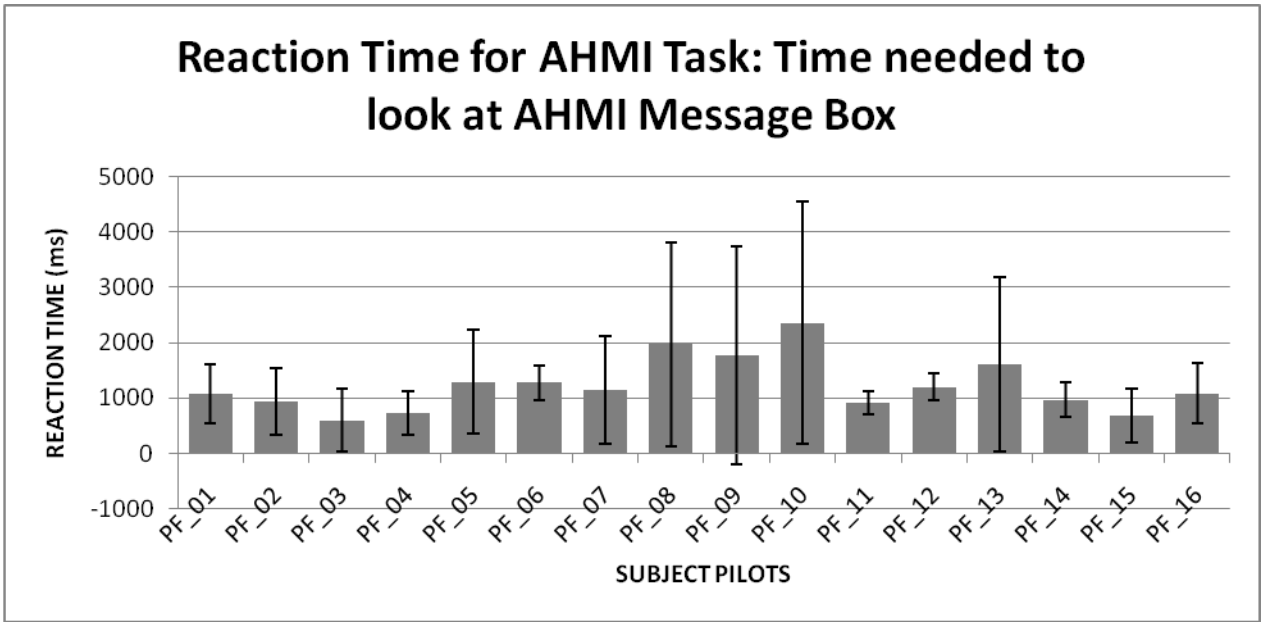

Figure 11: Reaction time of human pilots to the AHMI message box popup.

A comparison between model and human performance has been performed. The average reaction time of human pilots during cruise and approach phases is close to the time needed by the model. Results are depicted in Figure 12.

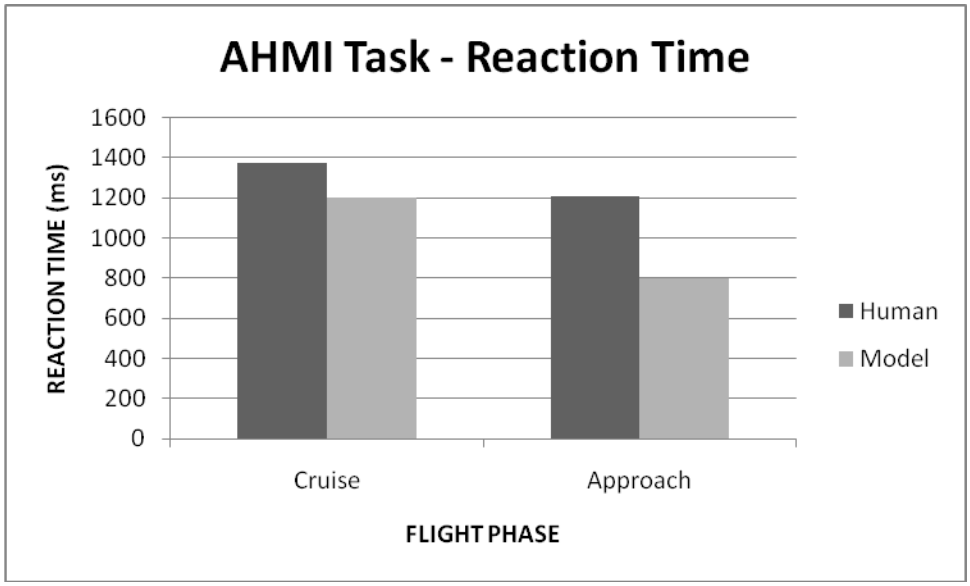

Figure 12: Average reaction time of human pilots and model to AHMI uplink event.

6.6 Analysis of H19

If pilots have reached a certain level of routine in a task the execution time of pilots within this task should be significantly similar for all pilots. In Figure 13 the task execution time of human pilots is depicted. Analyses have been conducted for the experiments in the Avionic Test Bed, where the AHMI procedure has been trained and for the experiments in the GECO. The error bars represent the individual standard deviation of each pilot. It can be seen that there is a difference between results of GECO and Avionic Test Bed. Currently, we assume that the differences can be explained by the more complex environment of the GECO (including more complex tasks). Results of the Avionic Test Bed demonstrate a much more consistent execution time for each pilot, where the results of GECO experiments are more distributed. In order to evaluate the reason for the high dispersion of performance data a context analysis of the concrete scenarios has to be performed for each pilot. In general, the results of the mean performance are within [42000;63000] ms for the GECO (PF_16 classified as outlier) and [23000;37000] ms for the Avionic Test Bed (PF_09 classified as outlier).

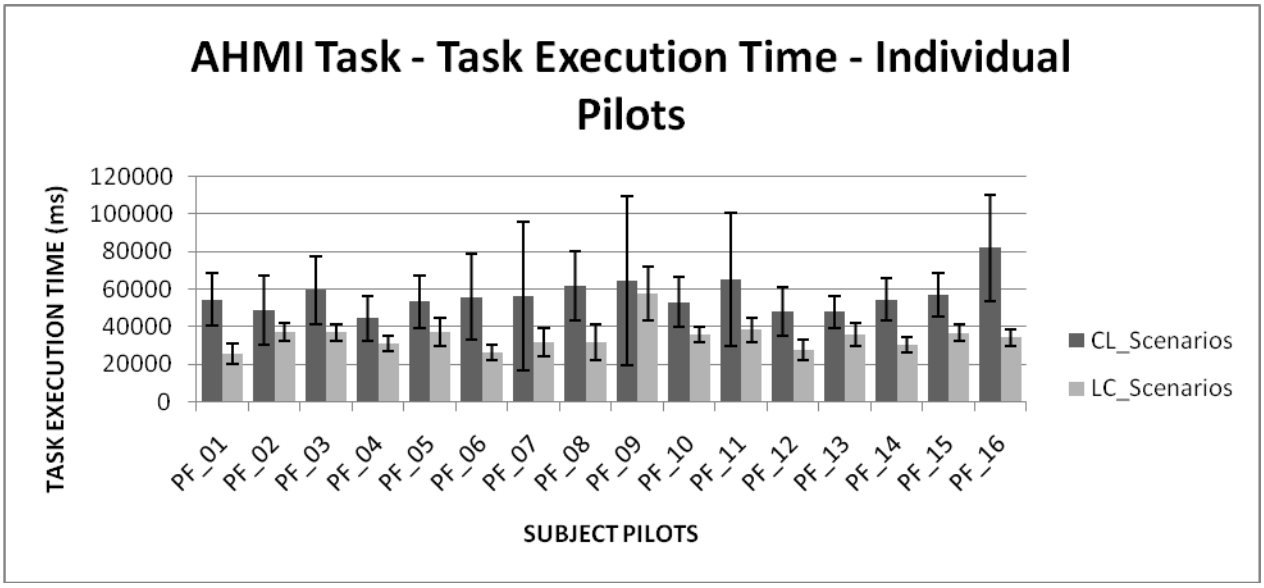

Figure 13: Task execution time of human pilots in dedicated AHMI task

We have compared the performance of human pilots with task execution time for the AHMI uplink task by the pilot model. The analysis shows that the model needs in average 25 seconds to complete the AHMI task (SD=4 seconds). These results significantly differ from human pilot performance in the same task. However, taking into account the differences between human pilots in the LC scenarios and the CL scenarios the reason for these differences may be seen in the basis, which is relevant for the model performance speed. We have taken the basic skill time provided by Fitt's law for motor actions such as using a mouse as device for moving a cursor on a user interface. The device used during the LC scenarios was a classical mouse device and pilots rapidly performed the task as described earlier. During the CL scenarios the device used was a trackball device which less easy to use and probably results in longer execution times for the task.

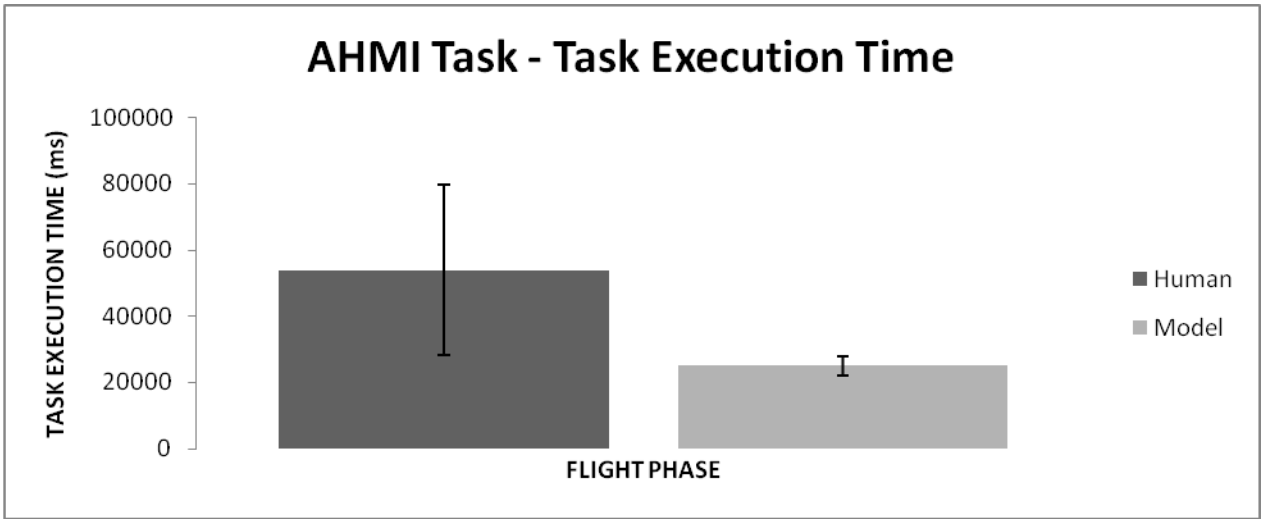

Figure 14: Task Execution Time of human pilots and the pilot model for the dedicated AHMI uplink task

6.7 Analysis of H21

A dedicated analysis of task execution time on the AHMI uplink task based on the performance data gathered during experiments in the Avionic Test Bed reveals that pilots learn to perform the task faster the more often they perform the task. In Figure 15 the task execution time of pilot in normal conditions (no errors injected) is presented. It can be seen that task execution time speeds up. In session 1 of experiments the task execution time has been measured with 45 seconds in average and with 30 seconds in average during session 5. This is an improvement of 33 percent.

The model does not improve execution time, because this kind of performance improvement is not part of the implementation of the model.

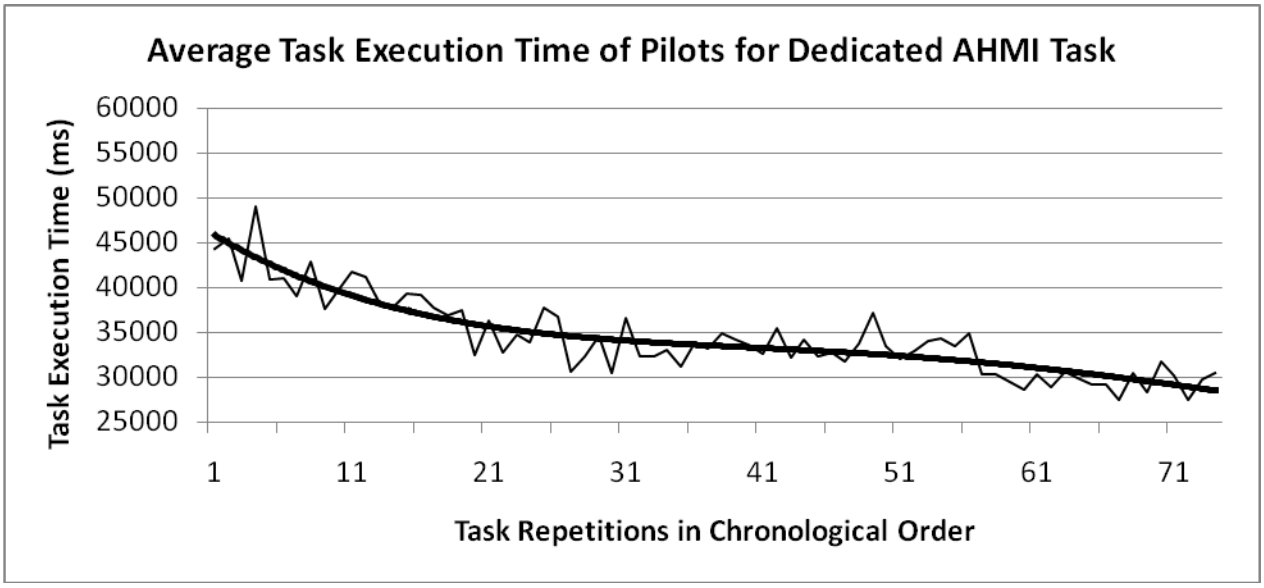

Figure 15: Task Execution Time of Pilots in Avionic test Bed Environment under normal conditions

6.8 Conclusion of data analysis for basic capabilities

The analysis of basic capabilities conducted during the second cycle shows that major improvements derived from the first analysis cycle have successfully been implemented. Especially those requirements concerning the visual performance of the model have been evaluated to be in line with human subjects' performance. Minor changes are necessary with regard to the speed of actions (resulting in a too fast task execution time). This issue is further discussed in section 8.1.

7 Analysis of Hypotheses about the Error production mechanisms

During cycle 2, the error production mechanisms Learned Carelessness, Selective Attention, and Cognitive Lockup have been analysed. The results are described in the next sections.

7.1 Data analysis Learned Carelessness

After improvements have been made with regard to the scenario design after cycle 1, we have tested human pilot's performance in a special experimental setup dedicated to analyse Learned Carelessness (LC). In contrast to cycle 1, experiments in cycle 2 have been conducted in a desktop workstation environment (called Avionic Test Bed) reducing the complexity of the full feature cockpit to the necessary displays. This decision has been made due to the fact that designing

experiments dedicated to the key variables was much easier. The experiments consisted of a primary and a secondary task. In the primary task pilots had to handle a flight plan uplink on the AHMI which was shown on an LCD-Display in front of the pilots. A mouse served as input device. In parallel to the primary task pilots had to handle the secondary task on a second LCD-display which was located in the left peripheral view with a second mouse as input device. The aim of the secondary task was producing a constant low level of workload as a supplement for monitoring activities which are performed under normal flight conditions. On the secondary task display pilots have been shown green or red rectangles in a random frequency and position on the screen. In case of a red rectangle appearing, pilots had to press the left button of the mouse and the right button in case of a green rectangle. Pilots had to press one of the buttons within 5 seconds otherwise the rectangle disappeared. Each erroneous or late selection has been counted. Pilots were asked to make as less errors as possible in order to provoke attention shifting between the primary and secondary task. Each subject pilot participated to 6 experiment sessions distributed over two days (day 1 = sessions 1-3, day 2 = sessions 4-6). Sessions 1 to 5 consisted of 30 scenario runs dedicated to build-up routine and LC. Session 6 consisted of 20 scenario runs dedicated to test LC. In each scenario run the pilots had to handle exactly one AHMI uplink resulting in about 170 datasets (few records failed due to technical problems). In total, 7 different types of scenarios have been designed. Scenario 0 represents a correct flight plan uplink. The scenarios 1-6 contain violations of exactly one check constraint. The scenarios have been defined as follows:

Scenario 0: Uplink of a correct flight plan
Scenario 1: Evaluation of c_1 = *false,* c_2-c_6 = *true*
Scenario 2: Evaluation of c_2 = *false,* c_1 and c_3-c_6 = *true*
Scenario 3: Evaluation of c_3 = *false,* c_1-c_2 and c_4-c_6 = *true*
Scenario 4: Evaluation of c_4 = *false,* c_1-c_3 and c_5-c_6 = *true*
Scenario 5: Evaluation of c_5 = *false,* c_1-c_4 and c_6 = *true*
Scenario 6: Evaluation of c_6 = *false,* c_1-c_5 = *true*

Parameters c_1-c_6 are normative checks on the AHMI. Each of these checks had to be performed by the pilots in order to correctly handle an ATC uplink. In the following, the checks are described:

c_1: On horizontal view: The first waypoint of the flight plan must be located in front of the aircraft, cf. **Fehler! Verweisquelle konnte nicht gefunden werden.Error! Reference source not found.** for an example where this constraint is violated. If the waypoint is behind, the FMS would calculate a trajectory that flies back to the first waypoint.

c_2: On horizontal view: The pilot has to check, if the flight plan ends on the runway, i.e. if the last waypoint is on the runway. When the check is successful, the PF can generate the trajectory by pressing "Load+Gen".

c_3: On horizontal view: Pilots have to check that the generated trajectory does not contain circles. This can happen, if too much time between the first check and the second check has passed by, and the aircraft has overflown the first waypoint before the trajectory is generated.

c_4: On vertical view: The pilot has to check, if the cruise flight level (CFL) is appropriate (in our experiments it should be over 8000 feet and below 32000 feet).

c_5: On vertical view: The pilot has to check that the interception[2] altitude is appropriate for the airport and runway.

c_6: On vertical view: The pilot has to check that the altitude of the last waypoint of the trajectory is equals the altitude of the runway.

It has been decided to prepare 2 different experiment setups. Setup A was dedicated to provoke LC and setup B served as control experiment. The design of experiments for setup A and setup B are depicted in Figure 16.

[2] The Intercept Altitude is the minimum altitude for intercepting the glideslope.

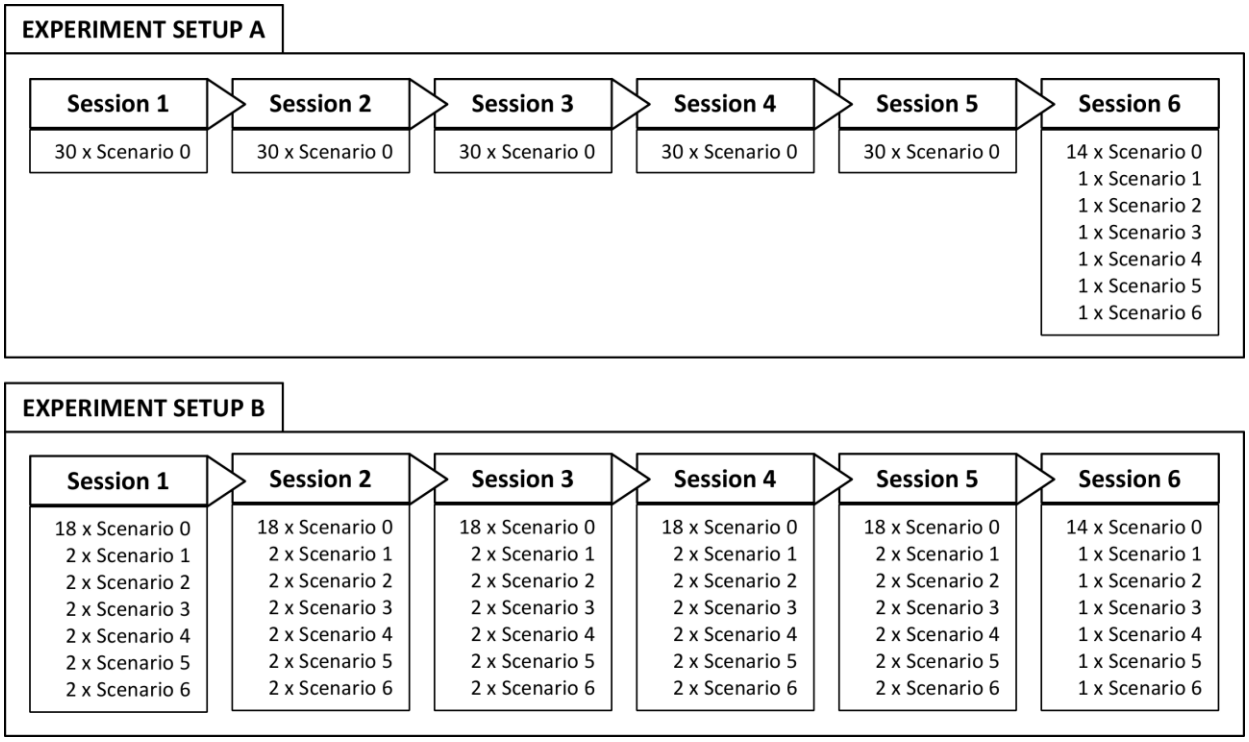

Figure 16: The design of experiments for testing Routine and Learned Carelessness was divided in two setups. Setup A was dedicated to provoke Learned Carelessness and setup B served as control experiment.

In setup A each of sessions 1-5 contained 30 instances of scenario 1. Session 6 contained 14 instances of scenario 1 mixed with each of scenarios 2-7. In setup B each of sessions 1-5 contained 18 instances of scenario 1 randomly mixed with each of scenarios 2-7 two times. Session 6 was designed similar to setup A. Thus, following the assumptions about LC presented in section 4, the effect of Learned Carelessness would be much more probable in setup A than in setup B. Output of these experiments are timestamp based datasets containing information about percept and motor actions, system and environment states.

7.1.1 Model data and comparison with simulator data

We have developed a systematic approach for validating our LC model based on the experiment data. The approach takes into account four indicators I_1-I_4 which can be positive or negative. Positive indicators confirm and negative indicators reject our LC model. Each indicator is based on one or more performance measures. In the following we will first present the performance measures used and second describe the indicators in detail.

- Task Execution Time ($m_{exe\_time\_task}$): Temporal duration of a certain data segment from t_{start} to t_{end}, where t_{start} is the time where the task has been started and t_{end} is the time where the task has been completed.
- Error Rate in scenario 0 per session ($m_{errors\_session}$): Relative number of incorrect procedure executions.
- Error Detection Rate in scenarios 1-6 per session ($m_{detection\_session}$): Relative number of detected flightplan errors.
- Error Detection Rate per check ($m_{detection\_check}$): Relative number of detected errors for pilots in experiment setup A.
- Susceptibility rating of each check ($m_{susceptibility\_check}$): Subjective rating of each check c_i with regard to parameters *risk* and *effort*.

For each measure we have prepared the data in data pre-processing steps. Each measuring point refers to aggregated pilot performance data. Results of $m_{exe\_time\_task}$ have been calculated for each experiment setup per run of scenario 0, which represent a full walk through the normative procedure under investigation. Because only runs of scenario 0 have been taken into account for $m_{exe\_time\_task}$ results of setup A contain more data points than results of setup B. Results are depicted in Figure 17. Bold lines are approximations to data, which are used to visualize trends. Results of $m_{errors\_session}$ are depicted in Figure 18. In Figure 19 the results of $m_{detection\_session}$ are depicted. Because experiment setup A contained incorrect flight plans only in session 6, only one data point exists for setup A. For setup B, we present results for all sessions. Results of $m_{detection\_check}$ represent the actual error detection rate of pilots in setup A per check c_i. In addition, each check c_i has been subjectively evaluated by pilots according to the parameters *effort* and *risk*. The evaluation has been conducted on a rating scale (n=6) reaching from 0 (low) to 5 (high). In order to evaluate the LC susceptibility of each check c_i, we aggregated the ratings of the two parameters for each check c_i with an equal weighting on the parameters *effort* and *risk* as shown in formula 1.

$$susceptibility_{c_i} = mean\left(\frac{effort_{c_i}}{n-1}, 1-\frac{risk_{c_i}}{n-1}\right) \quad (\mathbf{1})$$

The susceptibility ranges from 0 (low) to 1 (high). Both, the actual error detection rate and the LC susceptibility are depicted in Figure 20.

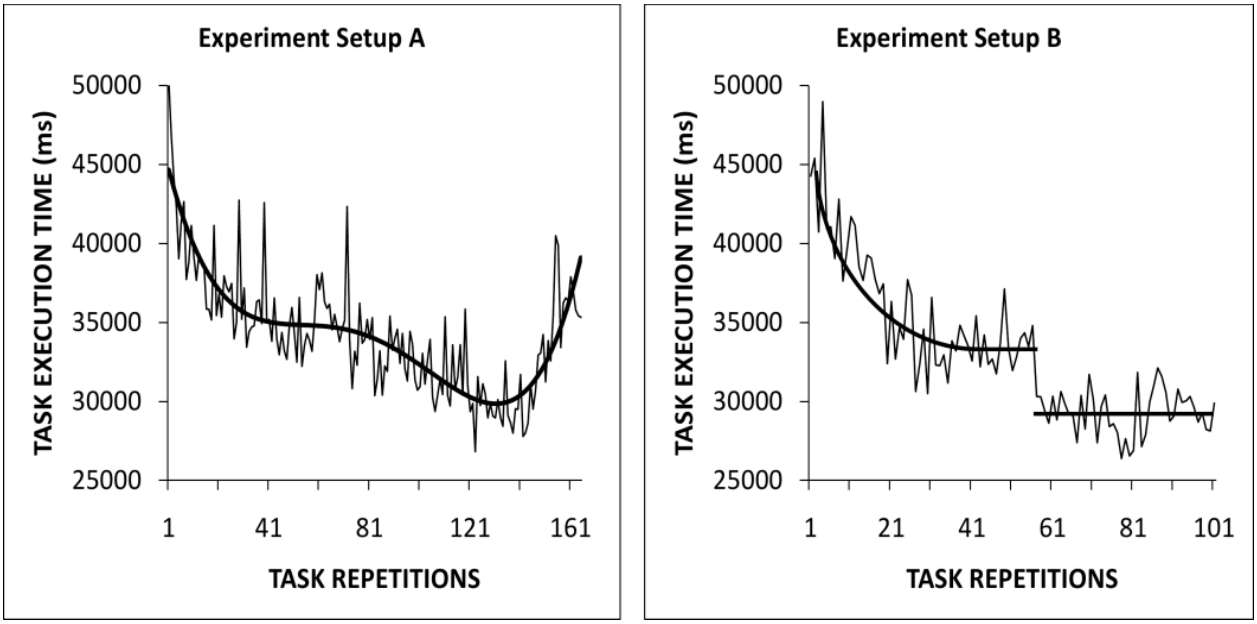

Figure 17: Task Execution Time of scenario type 1 AHMI uplinks in experiment setup A (left) and setup B (right) in chronological order ($m_{exe\_time\_task}$)

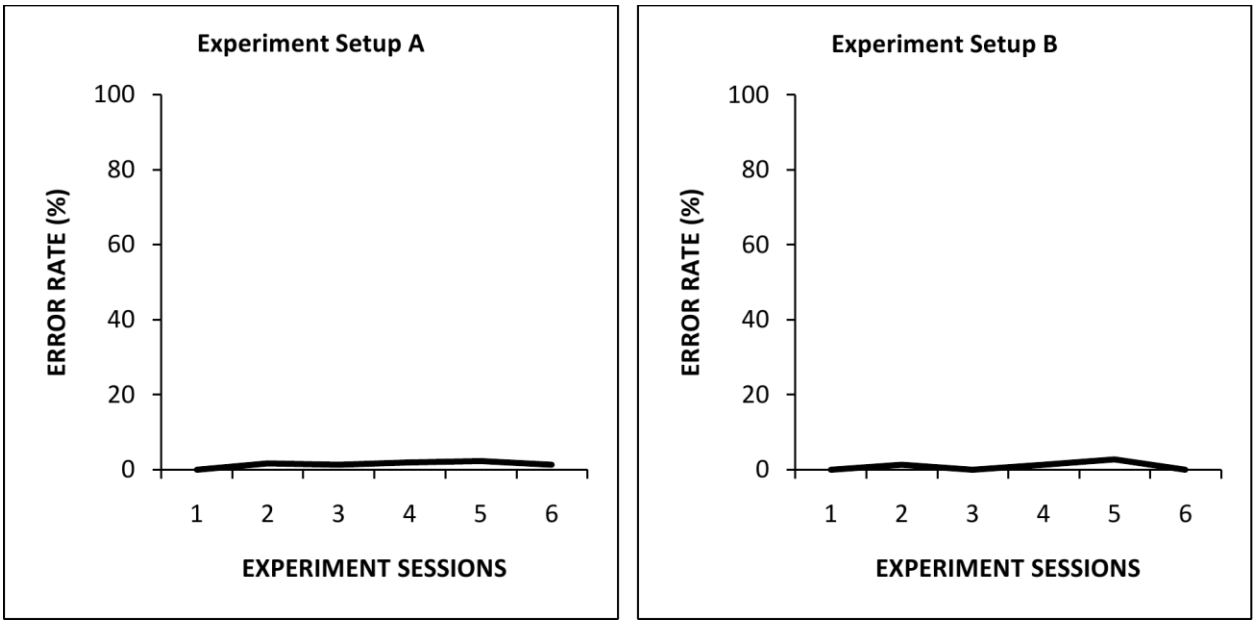

Figure 18: Error rates in experiment setup A (left) and setup B (right) for scenario type 0 per session ($m_{errors\_session}$)

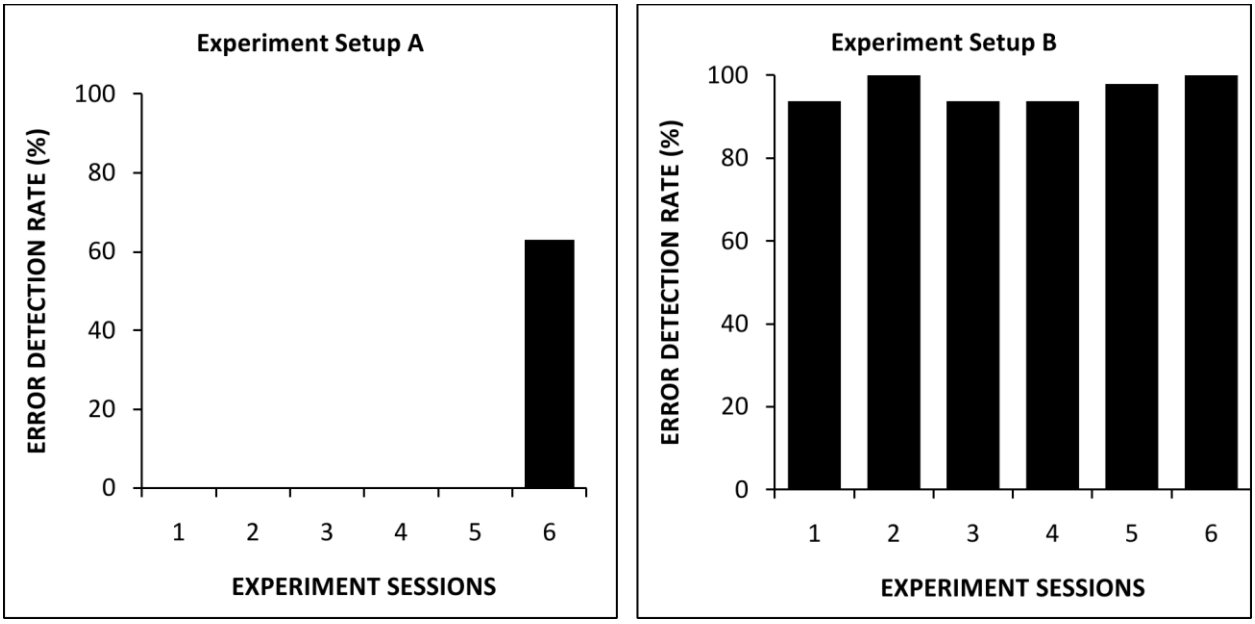

Figure 19: Error Detection Rates in experiment setup A (left) and experiment setup B (right) for scenario types 1-6 per session ($m_{detection\_session}$)

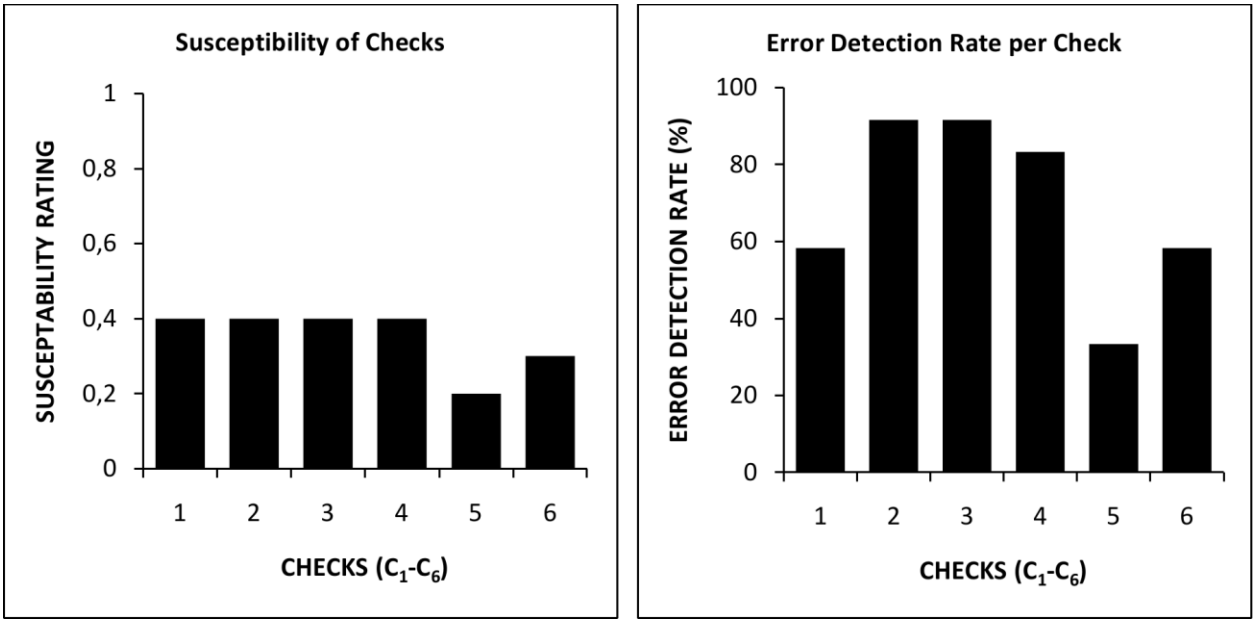

Figure 20: Susceptability of each check ci for Learned Carelessness ($m_{susceptability\_check}$) (left) and actual error detection rate per check ci for pilots of experiment setup A ($m_{detection\_check}$) (right)

After the performance measures used for our analyses are described, we will now focus on the indicators I_1-I_4:

I_1: The trace observed contains a data segment $s_{inductive}$, where the inductive learning phase is completed.

Relevant measures: $m_{exe\_time\_task}$, $m_{errors\_session}$

As described in section 3, LC emerges if the inductive learning phase has been completed. We assume that the traces observed contain a segment $s_{inductive}$ where the inductive learning phase has been completed. Segment $s_{inductive}$ can be identified by analyzing the task execution time and the error rate. The task execution time should converge to an optimum level of performance (cf. Figure 21). Further on, no procedure execution errors should disappear.

I_2: The trace observed contains a second data segment $s_{deductive}$, where certain procedure steps have been omitted.

Relevant measures: $m_{exe\_time\_task}$

We assume that LC emerges in a segment $s_{deductive}$, after the inductive learning phase has been completed. Because certain procedure steps will be omitted in $s_{deductive}$, we assume that the task execution time will decrease in this segment (cf. Figure 21).

I_3: Subjects make errors after LC has emerged.

Relevant measures: $m_{detection\_session}$

We assume that subjects do not detect incorrect flight plans after LC has emerged. This will be tested in experiment session 6 by analyzing the error detection rate.

I_4: Safety Precautions with high effort and low risk assessment are more susceptible for being omitted than others.

Relevant measures: $m_{detection\_check}$, $m_{susceptibility\_check}$

This indicator aim at validating the newly introduced parameters P_{risk} and P_{effort}. The result of measure $m_{detection\_check}$ will be compared to the result of measure $m_{susceptibility\_check}$. We assume a significant correlation between results of both parameters.

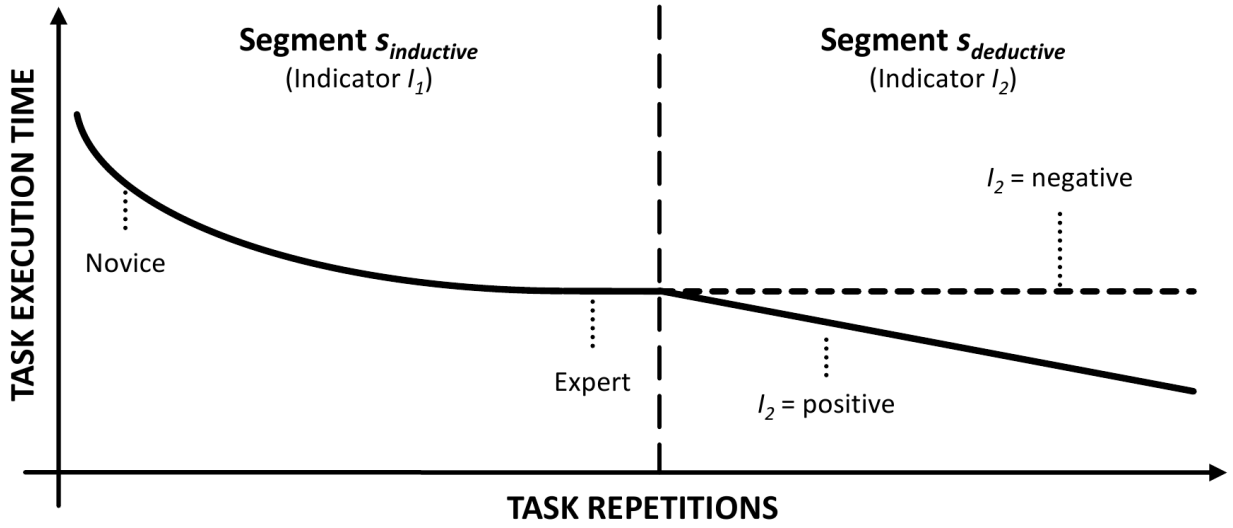

Figure 21: Indicators I_1 and I_2 describe the expected performance of subjects on parameter task execution time.

7.1.1.1 Analysis of I_1

According to our approach for analysis of I_1 we have taken into account $m_{exe\_time\_task}$ and $m_{errors\_session}$. After analysis of these performance measures we assume that pilots have completed the inductive learning phase within sessions 1-3. The average task execution time of the first 10 tasks has been measured with around 42000 ms in setup A and setup B. The analysis of $m_{exe\_time\_task}$ shows that the task execution time in both setups converges in session 3 around 35000 ms. This is an acceleration of 17%. The results of $m_{errors\_session}$ show that the error rate varying between 0% and 2.5% in setup A and between 0% and 3.5% in setup B. We have analysed each error in runs of scenario 0 in order to find out why and what kind of errors have been made by the pilots. An analysis of video material recorded during the experiments revealed that most of these errors can be attributed to data recording failures. Thus, we can assume that pilots have performed mostly correct.

7.1.1.2 Analysis of I_2

According to our hypothesis LC emerges after the inductive learning phase has been completed. Taking into account the experiment design, LC should emerge in sessions 4-5 of experiment setup A but not in setup B. Measure $m_{exe\_time\_task}$ has been selected for analysis of I_2. Results of $m_{exe\_time\_task}$ show that task execution time in sessions 4-5 of setup A speeds-up from 35000 ms to 29000 ms which is an acceleration of 11%. During sessions 4-5 of setup B no further speed-up has been measured. However, between sessions 3 and 4 (which are also separate the segments $s_{inductive}$ and $s_{deductive}$) a sudden decrease of execution time has been measured for experiment setup B from 35000 ms to 30000 ms. We assume that the reason is the break between day 1 and day 2. An explanation may be an increasing influence of fatigue during sessions 1-3. However, if fatigue is the reason of this phenomenon we cannot explain why the same phenomenon has not been measured in setup A experiments. Another explanation may be that pilots participating to experiments always expect errors during simulator flights, which may make them more careful. As a consequence, results of $m_{exe\_time\_task}$ reach a mean performance of 29000 ms in setup A and 30000 ms in setup B, which is very equal.

7.1.1.3 Analysis of I_3

According to this indicator we expected that pilots would not detect incorrect flight plans after LC has emerged. Results of session 6 will be used to test if this assumption holds. We have checked pilot actions of each run in session 6 against a formal model of normative procedures. All results have been aggregated and are represented by $m_{detection\_session}$. We have already explained that the results of sessions 1-5 of setup A contain only runs of scenario 0, thus there were no incorrect flight plans to detect. However, we analysed the error detection rate of pilots in setup B for all sessions in order to use the results for a comparison between setup A and setup B. The error detection rate in sessions 1-6 of pilots in setup B varies between 90% and 100% which means that pilots in setup B have been able to detect almost all incorrect flight plans. In contrast, results of setup A session 6 shows that the error detection rate is at 63%. This is a significant difference to all data points in setup B.

A side effect of injecting erroneous flight plans in setup A can be observed in Figure 17 which is a strong slow-down of task execution time together with an increase of actions performed. We assume that this effect can be explained by recovering of omitted actions and uncertainty when pilots became aware of their errors.

7.1.1.4 Analysis of I_4

This indicator will be used to analyse the newly introduced parameters *risk* and *effort* which we suspect to take influence of LC. We decided to test parameter I_4 based on $m_{detection\_check}$ and $m_{susceptibility\_check}$. According to our assumptions, the error detection rate of errors related to a check c_i should be low if the susceptability of c_i is high. We have performed a correlation analysis between susceptability and error detection rate based on Kendall's tau coefficient. The result ($\tau = 0.67$) shows a

relevant positive correlation between the parameters susceptability and error detection rate. Calculation of the correlation between risk and error detection rate (τ = -0.65) on the one hand, and effort and error rate (τ = 0.61) on the other hand separately revealed a direction of correlation which is contrary to our assumptions. According to these findings, LC would emerge if risk is high and effort is low. These results lead us to think about three possible explanations for the findings: (1) The subjective rating by pilots was not correct, (2) the parameters risk and effort are not sufficient to model LC, or (3) the findings are actually correct. In order to evaluate these different options, further analyses are necessary. Thus, a final evaluation of the parameters is currently not possible.

7.1.1.5 Comparison

Our model of LC has systematically been analyzed based on the indicators I_1-I_4. The analyses of these indicators show differentiated results. We have been able to define a segment of data $s_{inductive}$ where we are confident that the inductive learning phase has been completed (I_1). We have also been able to find a segment $s_{deductive}$ in setup A where we assume that LC has emerged (I_2). We have detected a sudden decrease of execution time between $s_{inductive}$ and $s_{deductive}$ in setup B. Thus, pilots in setup A did not perform faster than pilots in setup B. Further analyses are necessary to find an explanation for this phenomenon. In session 6 of setup A pilots made much more errors than subjects in setup B. We concluded that subjects in setup A omitted relevant checks which lead to execution of an inadequate procedure path (I_3). Finally, we were not able to show a relation between the error rate of specific errors and the parameters risk and effort as we have assumed. In contrast, the results showed a correlation in the other direction, meaning that LC is probable to occur if the parameter effort is low and the parameter risk is high. We do currently not have an adequate explanation for these results. Thus, further analyses are required here as well. All results of our analyses are shown in Table 4.

	I_1	I_2	I_3	I_4
Setup A	positive	positive	positive	*
Setup B	positive	negative*	negative	*

Table 4: Results of Indicators I1-I4 for experiment setups A and B, (*) indicates that further analyses are required

Although we have been able to identify certain effects which can be caused by LC, a problem of experiment design for testing LC is the factor of time. LC is an effect, which can be observed under normal conditions after a long period of learning. Our experiments have been conducted in two days full of repetitive tasks. The question is, if this kind of time-lapsed experiment affects the results in a way that the effects observed are caused by other mechanisms, such as fatigue, monotony, priming or mental saturation. However, the potential benefit for system designers of modelling

error producing mechanisms such as LC and simulating them within a cognitive architecture legitimates for further investigation.

7.2 Data analysis Selective Attention

7.2.1 Simulator data

Selective Attention (SA) has been analysed based on the visual events emerging in the visual field of the human pilots. Normally, the visual event triggers a reactive shift of attention towards the area of the visual event. If the visual attention is focussed on another area, it might happen that a visual event (even if it is in the visual field) goes undetected.

Our experiments focussed on the flight mode annunciations (FMAs) on the PFD. During each simulator flight, a mode change is visually signalled by a flashing on the FMAs. We assumed that pilots will be triggered to look on the FMAs. A similar analysis has been conducted by Mumaw et al. (see [1]).

A total of 1924 events has been analysed for the human pilots experiments. Pilots' visual focus has shifted to the FMAs during the first 10 seconds (the flashing disappeared after 10 seconds) in 19.5% of the cases. Further on, we tested if pilots would react to the event within the next 10 seconds, where the flashing has already disappeared, because we assumed that some pilots may have recognized the event peripheral but shifted the task to look at the changes because their current task had a higher priority. Within the first 20 seconds 28% of all events have been reacted to by a visual focus on the FMAs. These results are generally in line with the results reported by Mumaw et. al. in [1].

7.2.2 Model data and comparison with simulator data

The model experiments conducted contained a total of 229 visual events. 39% of these events have been followed by a shift of visual focus to the FMAs within the first 10 seconds. The larger interval (≥ 20s) has not been considered, because reactions after the flashing (due to shifting of task) have not been modelled. Results for human pilots, the model and literature results by Mumaw et al. are depicted in Figure 22.

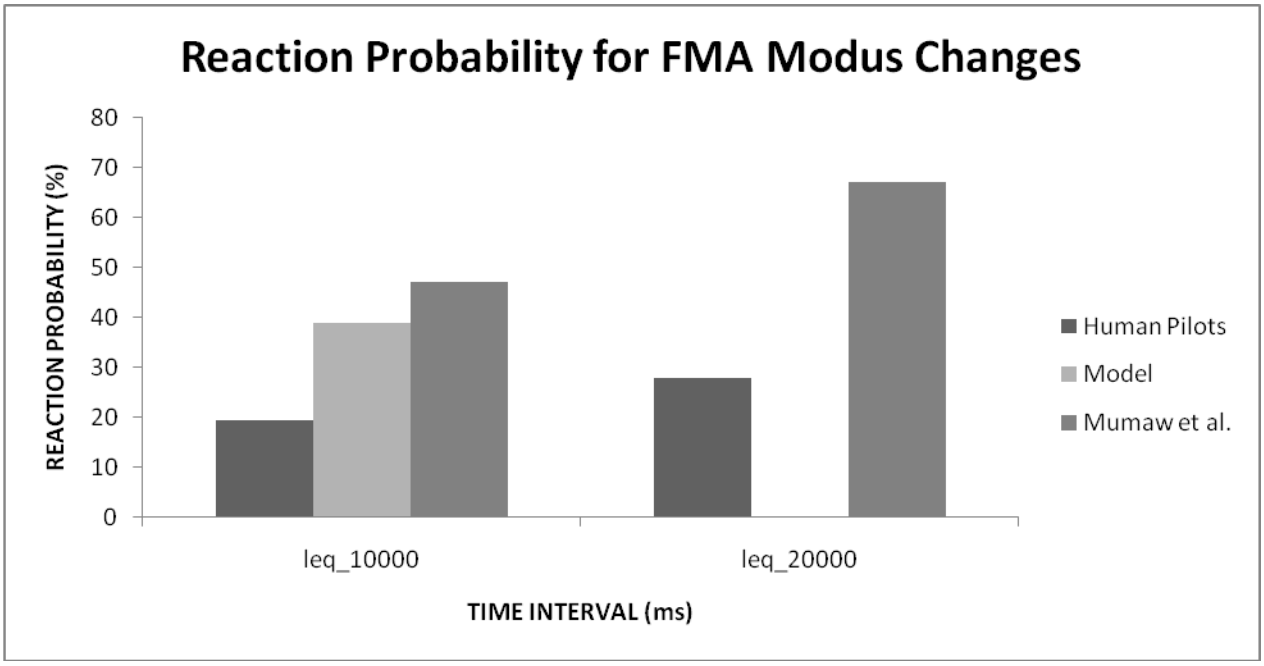

Figure 22: Reaction Probability for FMA Modus Changes

7.3 Analysis of H32, the subjective experience of cognitive lockup scenarios

The hypothesis that is evaluated in this section is the following: At moments in the scenario where high workload is expected, the pilots subjectively experience a high workload.
This hypothesis is very relevant to the evaluation of cognitive lockup. As described in D4.6, we hypothesise that the workload a person experiences is directly linked to the occurrence of cognitive lockup. We base this hypothesis on the Cognitive task load theory of Neerincx [2], in which the parameters task set switches (TSS), time occupied (TOC), and the level of information processing are underlying concepts of workload. For a more detailed description of the theory and the setup of the scenarios for this experiment, please see D4.6.

Using knowledge on the aviation domain, general human factor knowledge and the cockpit technology, cognitive lockup was chosen as an EPM with a high chance to occur (see D1.2). The scenarios were built up using the human factor knowledge on cognitive lockup. As mentioned above, the following two parameters were considered as drivers for cognitive lockup:

- Task switch sets (TSS)
- Time occupied (TOC)

Based on these two parameters 4 scenarios were developed and 3 control scenarios. The control scenario for scenario 1 and 4 was the same (see D4.6).

Two working hypotheses were derived:

1. If the TSS is "high" and the TOC is "high" the switch to another (higher priority) task is not done on time.
2. If the TSS is "high" and the TOC is "high" the experienced effort is "high" then the reaction time to switch to another (higher priority) task than in situations where TSS and TOC are lower.

As said above, we hypothesised that when the TSS and TOC are "high", the effort is "high" as well. This is also evaluated during the experiments.

7.3.1 Simulator data for hypothesis 32

During the experiment, the pilots filled in questions on experienced effort (on the RSME scale [3]), experienced time pressure (TOC) and whether they felt there were many tasks to attend to (TSS). Next to this, the experienced effort for selected times in the scenarios was asked.

We will first look at the general perceived effort, TOC and TSS by the pilots flying, comparing the control and test scenario.
For scenario 2 we exclude pilot 16 for TOC and TSS, because these were filled in on the wrong scale by the pilot during the control scenario. Pilot 9 is excluded for scenario 3, because the pilot did not fill in the answers for control scenario 3. The Likert scale questions were not filled in correctly by all pilots (choosing a point in between two options). We tested if the results were really different when choosing to round up (3.2 becomes 4) or down (3.6 becomes 3) or use normal round (3.4 becomes 3 and 3.5 becomes 4). This was not the case, so we used normal round.

7.3.1.1 Scenario 1

In test scenario 1 the pilot had to deal with an autopilot failure, a thunderstorm, anti-skid failure (need of longer runway than normally) and a runway change. The control scenario differed on one point; there was no autopilot failure. For a more detailed description of the scenarios, together with a motivation, please see D4.6.

Results show that in the scenario in which the autopilot failed, the pilot did not only objectively have more tasks, but the pilot also perceived the scenario as containing more tasks (Figure 23). Next to this the perceived TOC increased (Figure 24) and the effort was rated as significantly higher for the test scenario in comparison to the control scenario (Figure 25).

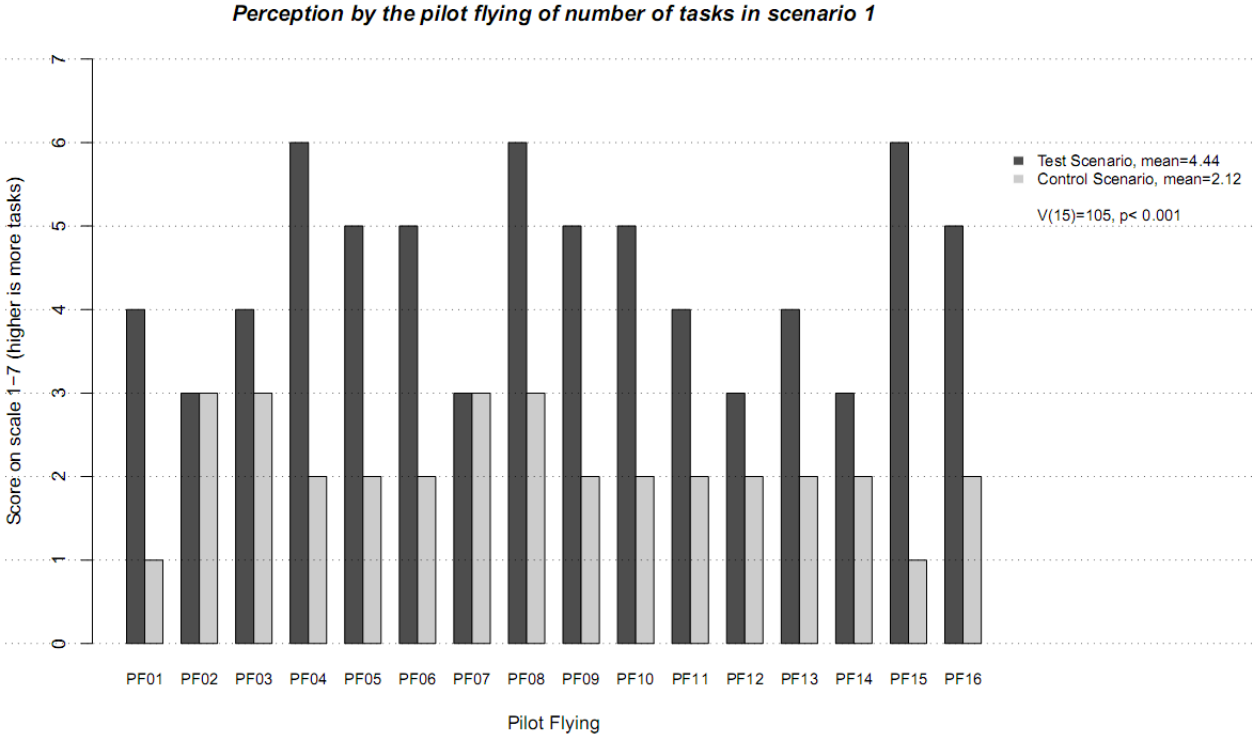

Figure 23: Perception by the pilot flying of the number of tasks in scenario 1.

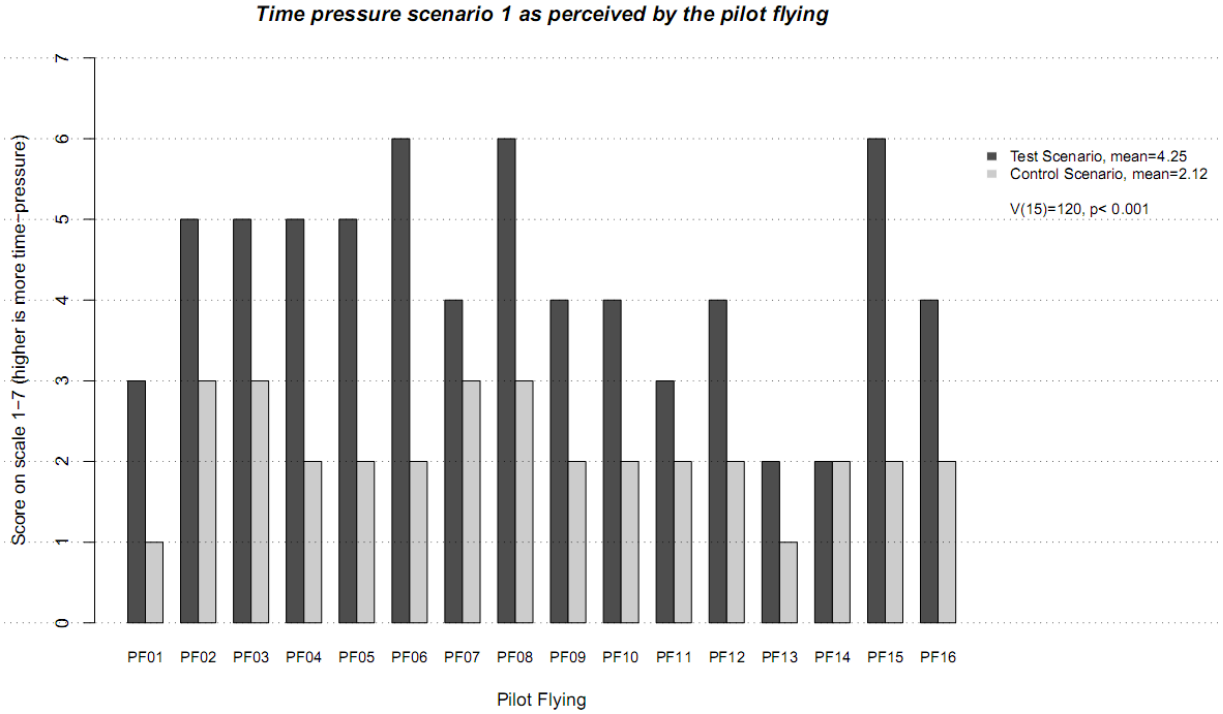

Figure 24: The time pressure in scenario 1 as perceived by the pilot flying.

In the first scenario we also looked at how the pilot monitoring rated the effort for the runway change of the pilot flying comparing the test and the control scenario (Figure 26). We also asked the pilot flying about the effort for the runway change in test and control scenario (Figure 25). Both rated the effort for the runway change in the test and control scenario as significantly different – the effort for the runway change was rated as significantly higher in the test scenario than in the control scenario. We excluded pilot 1 and 2 because they did not answer the questions so we could not compare them to the data by the pilot monitoring.
These results show that for scenario 1, the envisioned manipulation of the variables *number of tasks* and *time occupied* has been successful. Both variables have been experienced as significantly higher in the test scenario than in the control scenario. The same holds for the workload; the pilot flying had a significantly higher workload in the test scenario compared to the control scenario.
This is important
The scenario can be evaluated for cognitive lockup, which will be done in section 7.3.2.

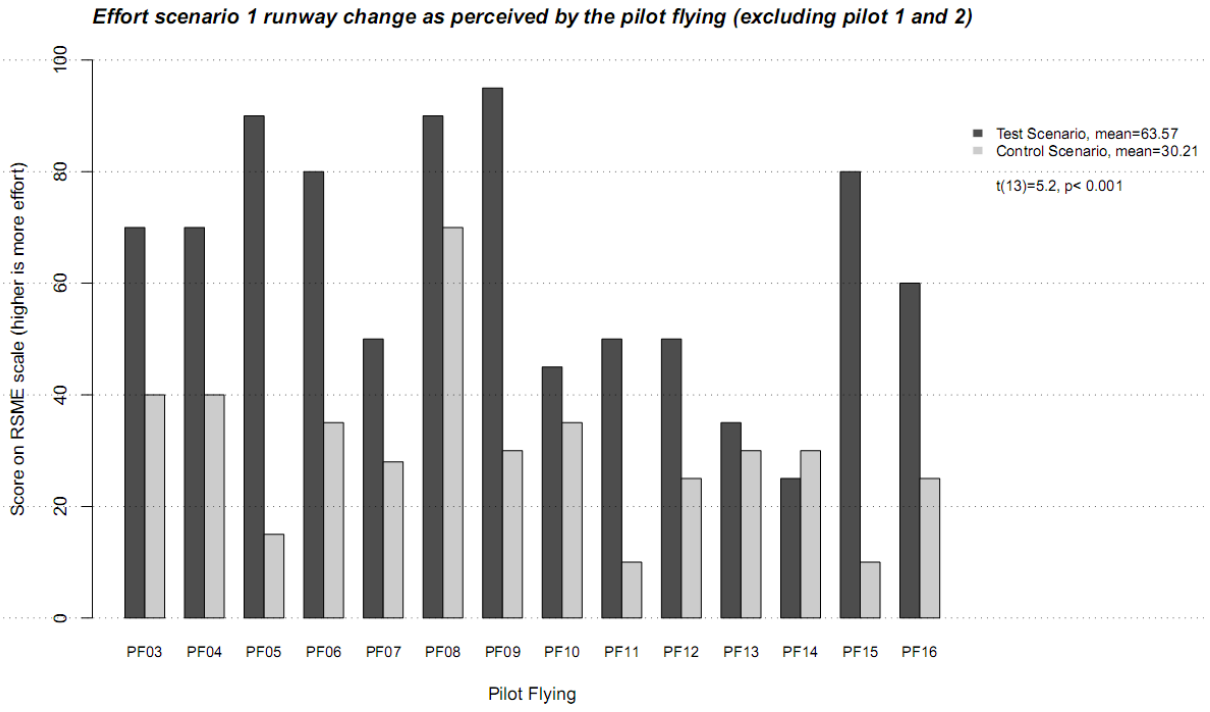

Figure 25: The effort of scenario 1 during the runway change as perceived by the pilot flying (excluding pilot 1 and 2).

7.3.1.2 Scenario 2

In test scenario 2 the pilot had to deal with an autopilot failure and a runway change late in the approach phase. The control scenario differed on one point; in the control scenario the runway change was earlier in the scenario.

Results show that the pilots did not experience the TSS (Figure 27) as significantly different between the test and control scenario. The TOC (Figure 28) was neither experienced as being significantly different between the test and control and the same held for the effort (Figure 29).

The effort for the runway change is rated significantly higher for the test scenario than for the control scenario by the pilots flying (Figure 29). The pilots monitoring do not rate this as significantly different (Figure 30).

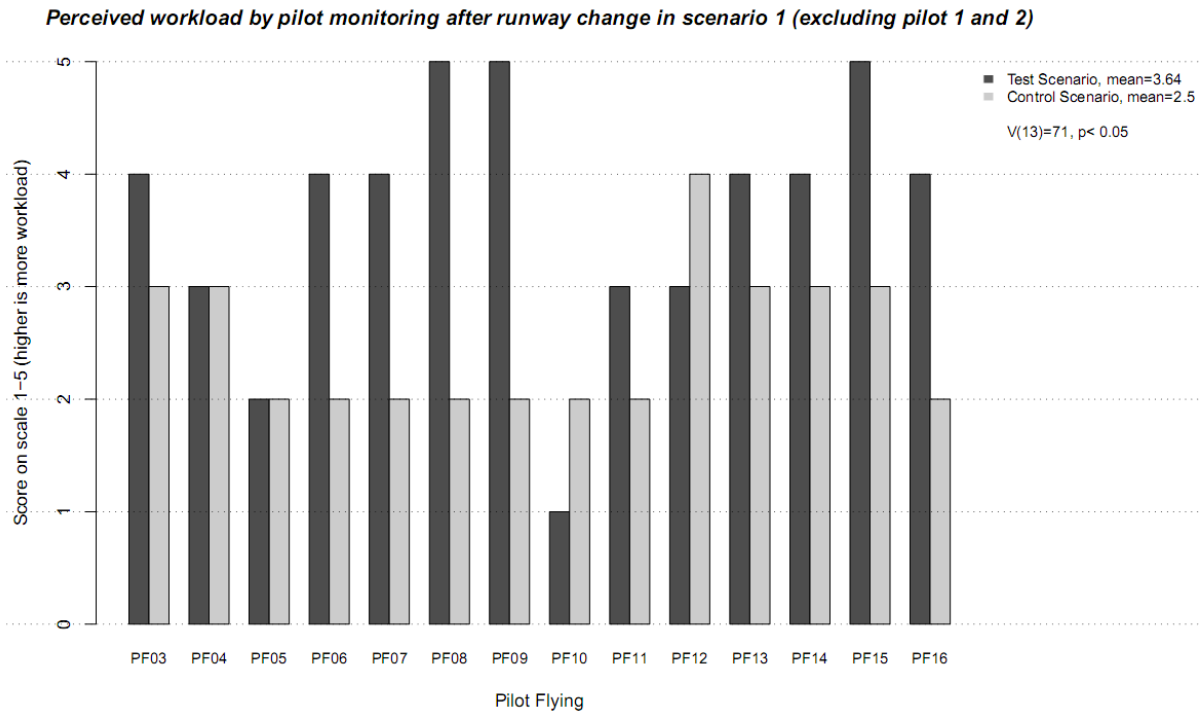

Figure 26: The workload of the pilot flying as perceived by the pilot monitoring after the runway change in scenario 1 (excluding pilot 1 and 2).

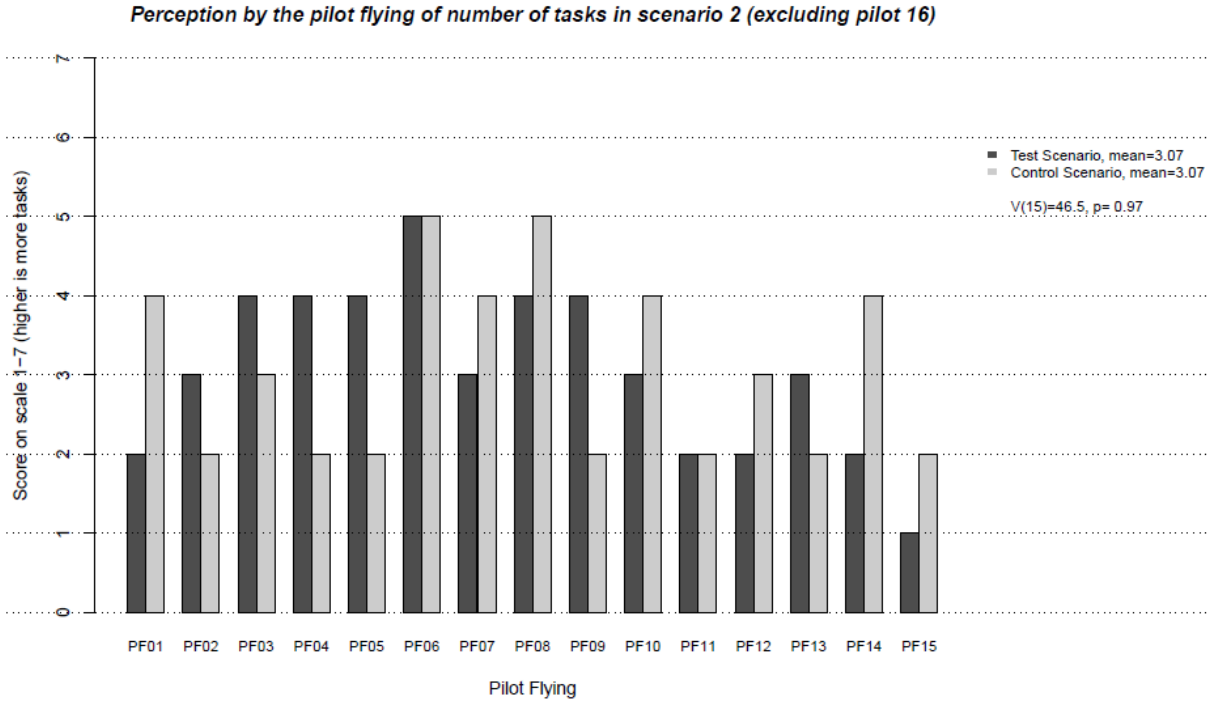

Figure 27: The perception by the pilot flying of the number of tasks in scenario 1 (excluding pilot 16)

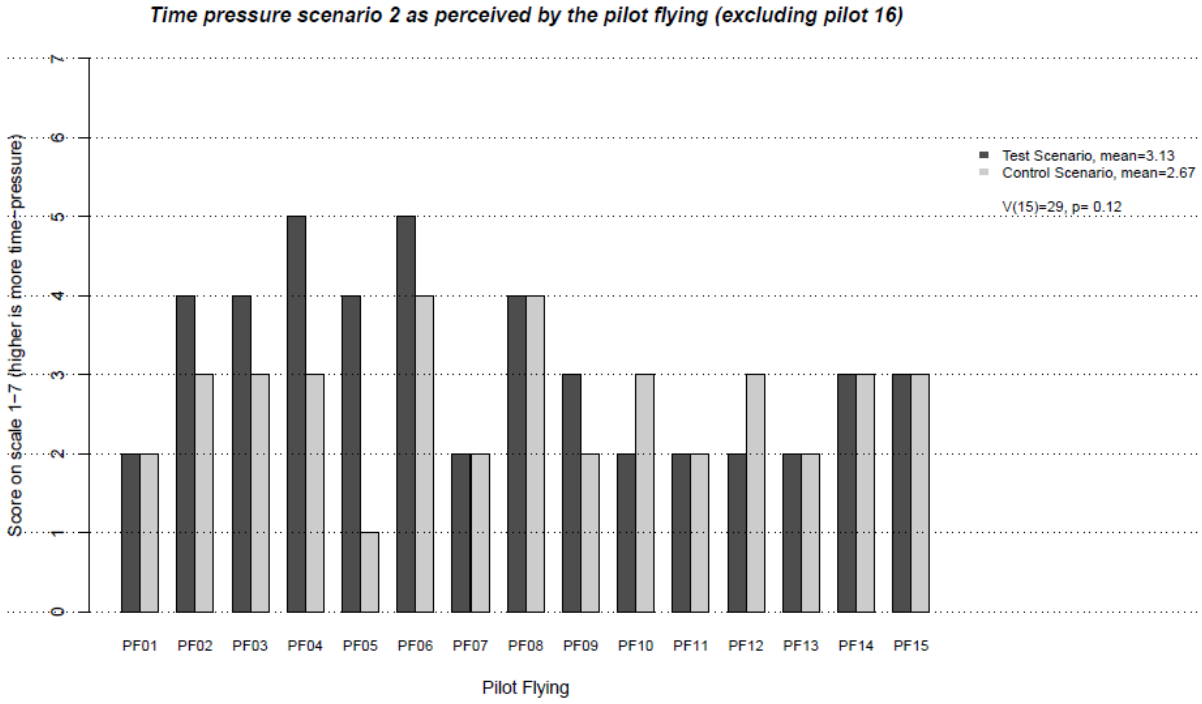

Figure 28: The time pressure in scenario 2 as perceived by the pilot flying (excluding pilot 16)

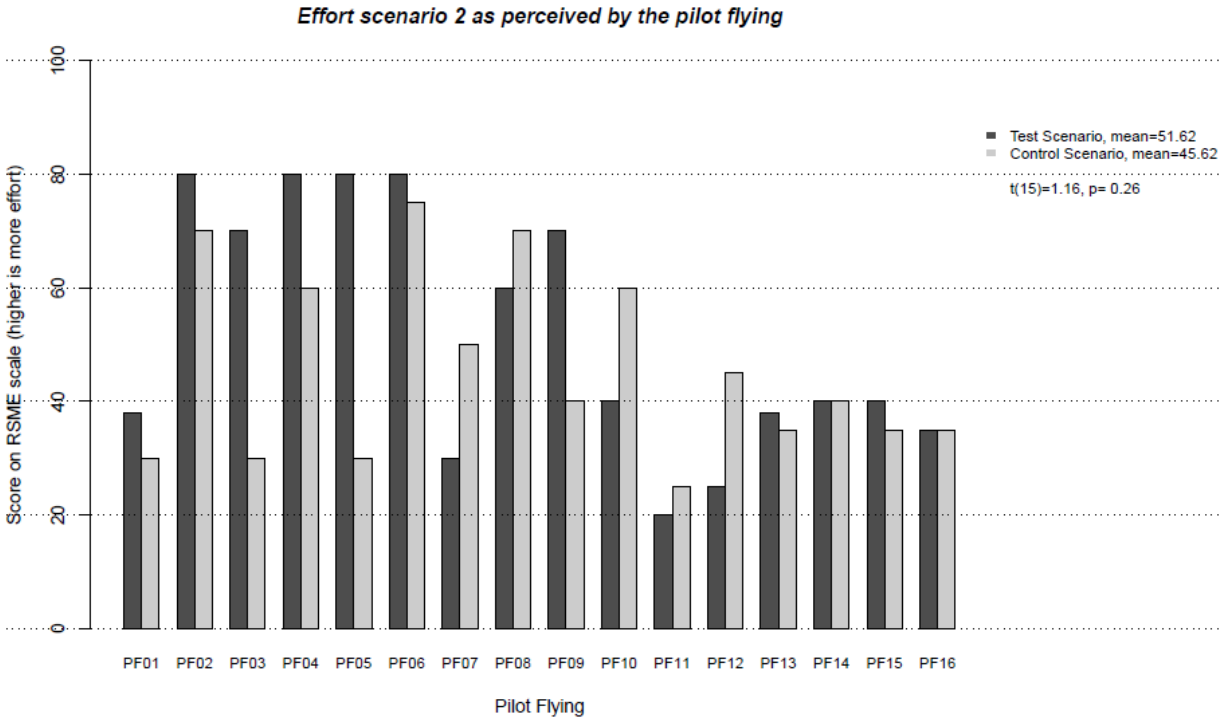

Figure 29: The effort in scenario 2 as perceived by the pilot flying

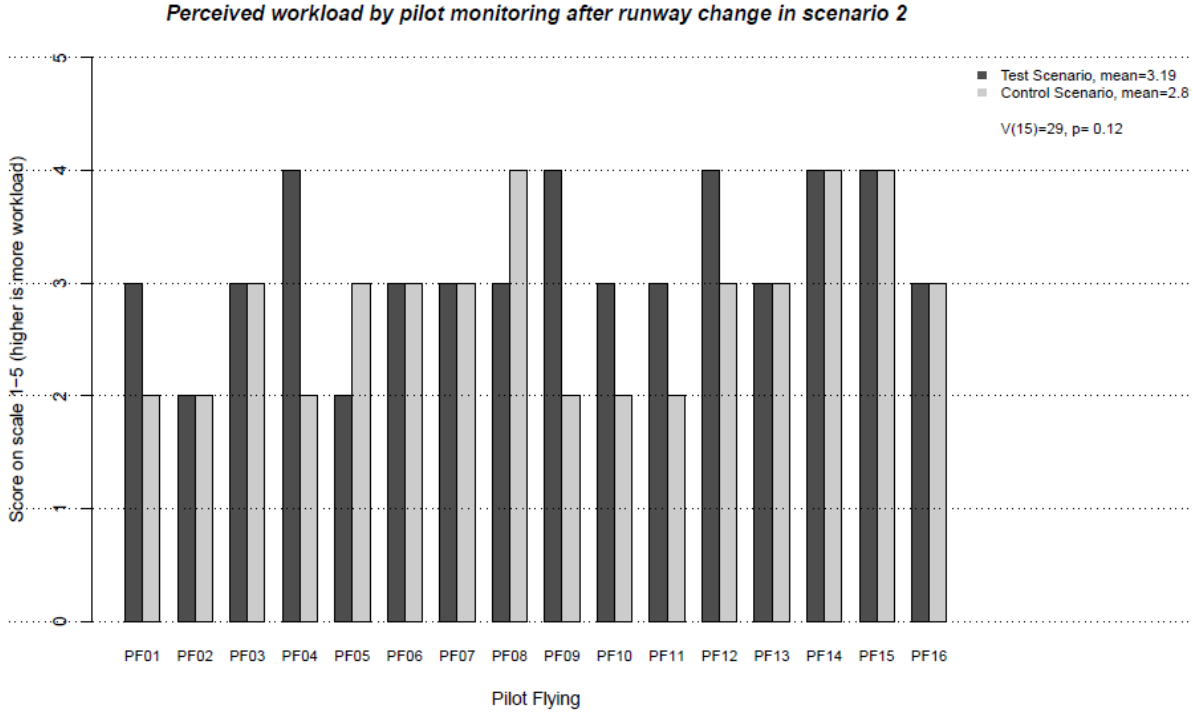

Figure 30: The workload of the pilot flying as perceived by the pilot monitoring after the runway change in scenario 2.

7.3.1.3 Scenario 3

In test scenario 3 the pilot had to deal with radar vectoring and a high intercept runway change while there are speed and height constraints. The control scenario differed on one point; in the control scenario the runway change was earlier in the scenario.

Results show that the pilots did experience the TSS (Figure 31) as significantly higher in the test than in the control scenario. The TOC (Figure 32) was not experienced as being significantly different between the test and control and the same held for the effort (Figure 33).

The pilot monitoring rated the effort of the pilot flying as significantly different in the test and in the control scenario (Figure 34).

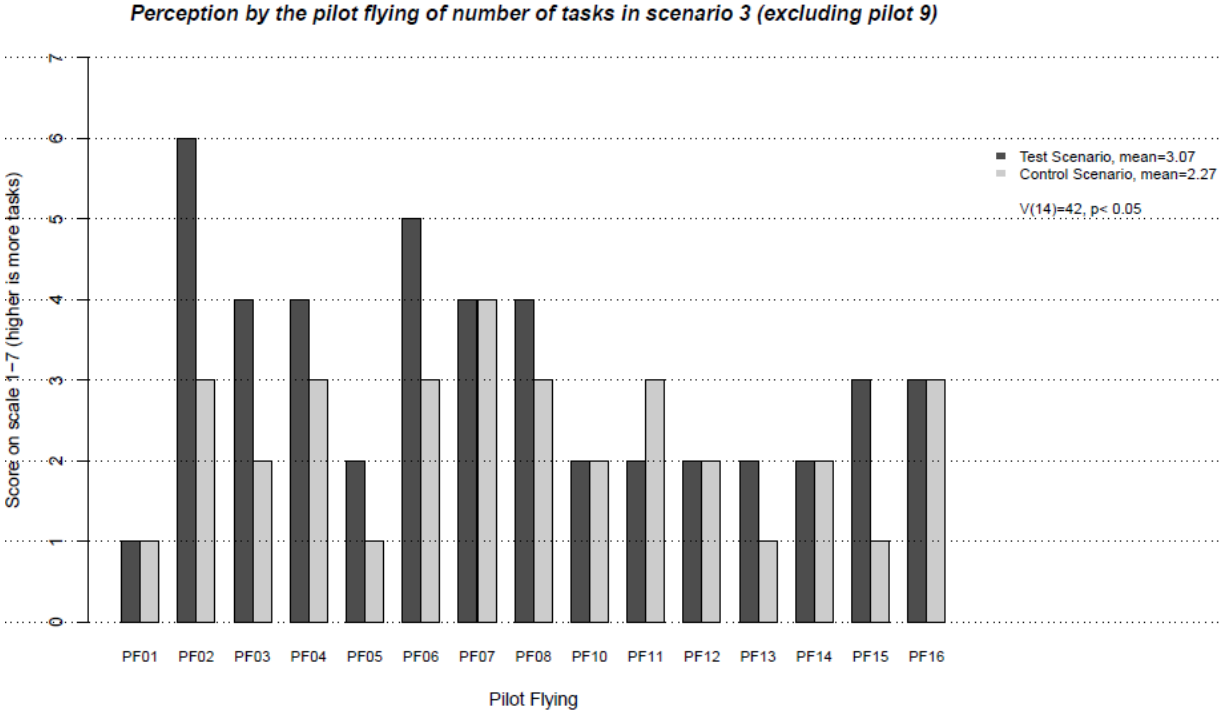

Figure 31: The perception by the pilot flying of the number of tasks in scenario 3 (excluding pilot 9).

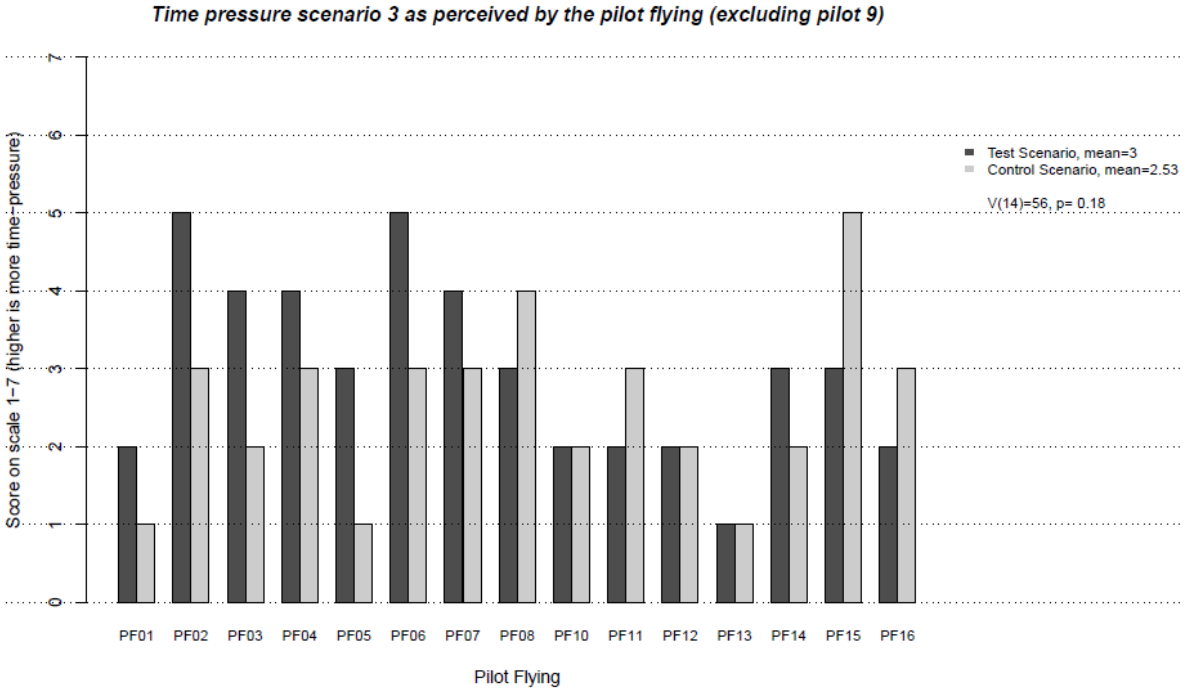

Figure 32: The time pressure in scenario 3 as perceived by the pilot flying (excluding pilot 9).

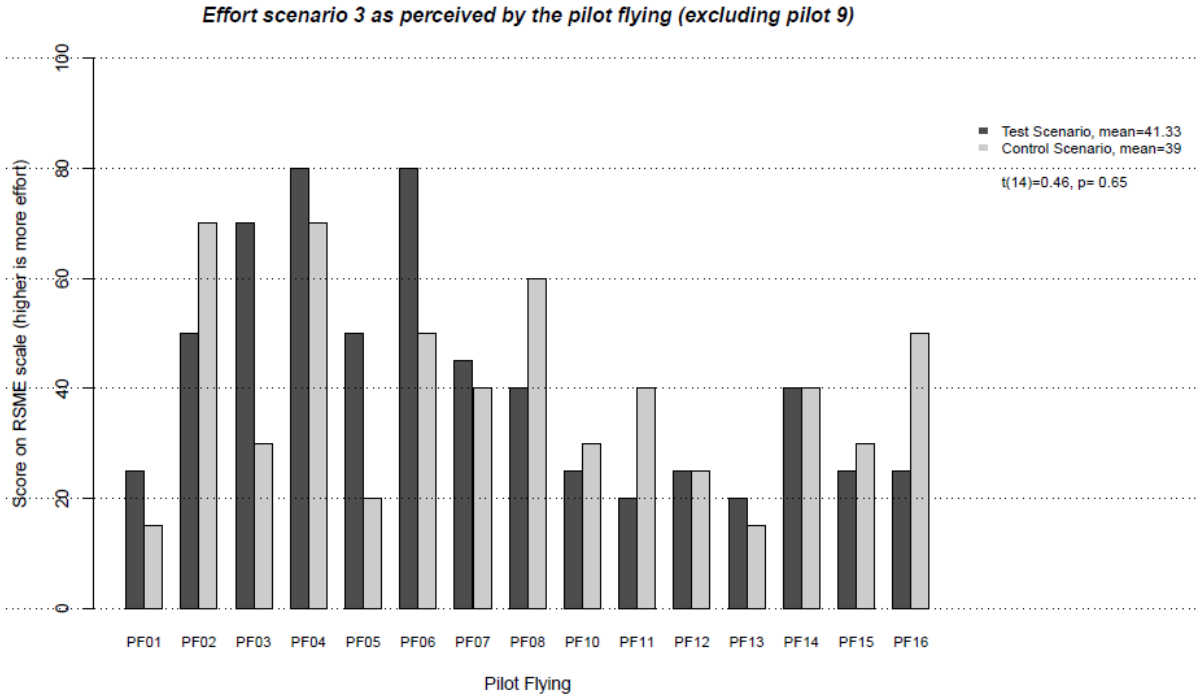

Figure 33: The effort in scenario 3 as perceived by the pilot flying (excluding pilot 9).

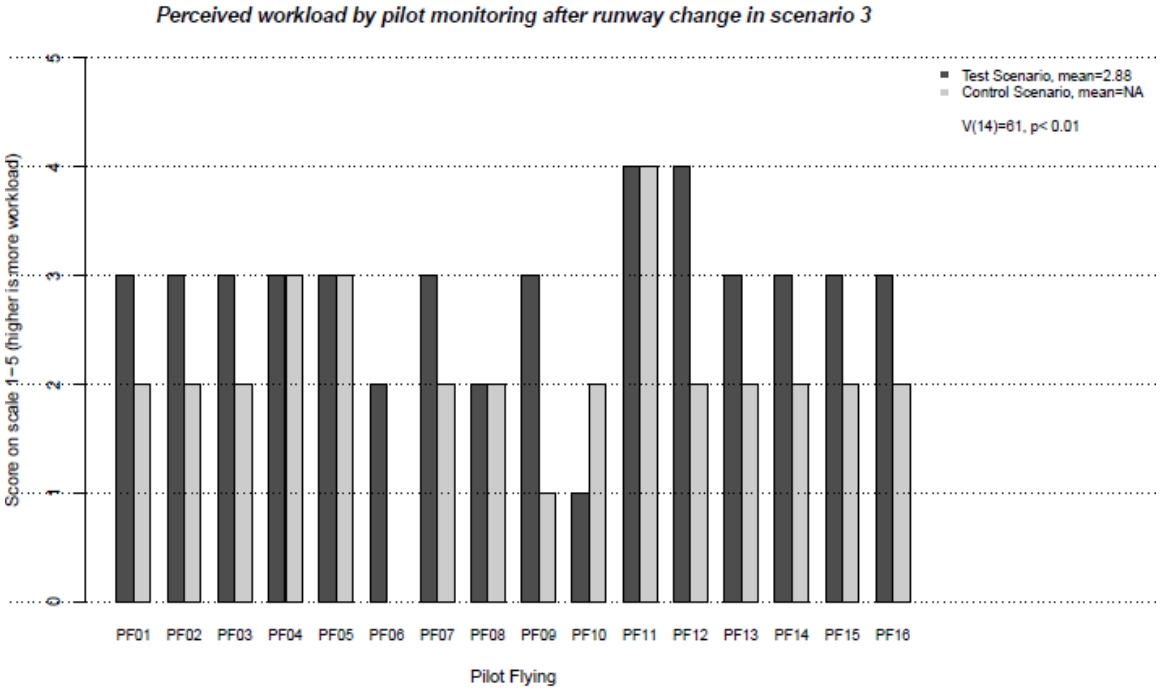

Figure 34: The workload of the pilot flying as perceived by the pilot monitoring after the runway change in scenario 3.

7.3.1.4 Scenario 4

In test scenario 4 the pilot had to deal with inputting waypoints and when he nearly finishes this there is a fuel pump malfunction. The control scenario was the same as for scenario 1.

Results show that the pilots did experience the TSS (Figure 35) as significantly higher in the test as in the control scenario. The TOC (Figure 36) and effort (Figure 37) were also experienced as being significantly different between the test and control. The pilot monitoring also rated the effort of the pilot flying in the test scenario as significantly higher than in the control scenario (Figure 38).

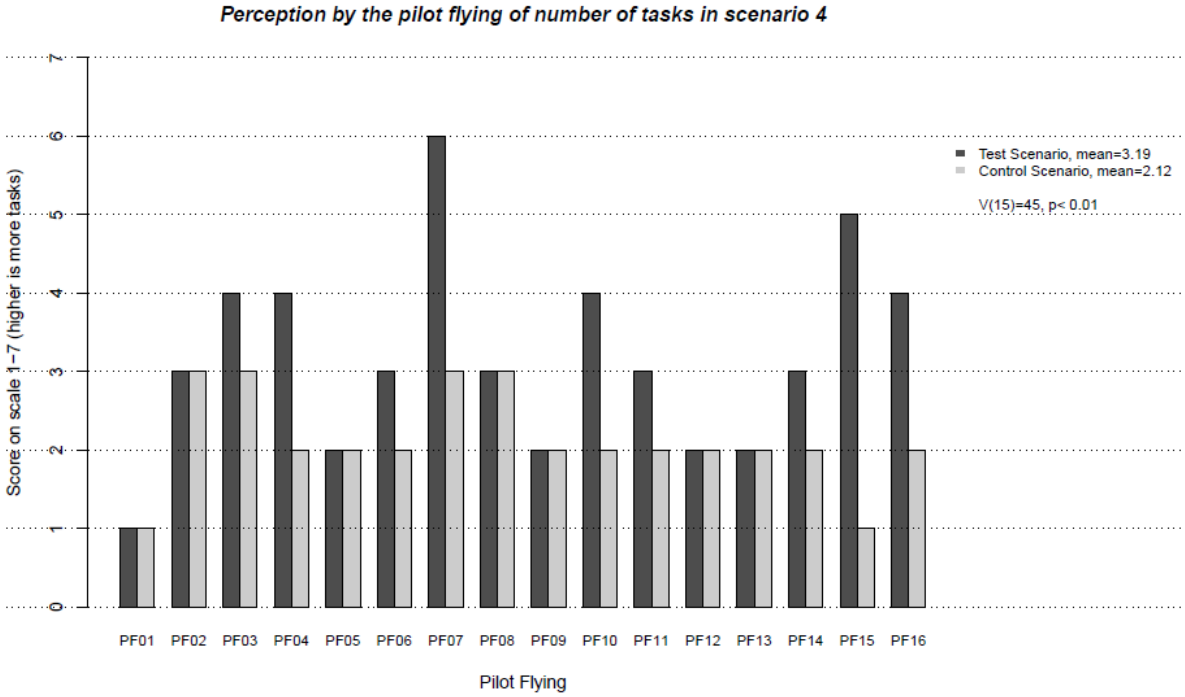

Figure 35: The perception of the number of tasks by the pilot flying in scenario 4.

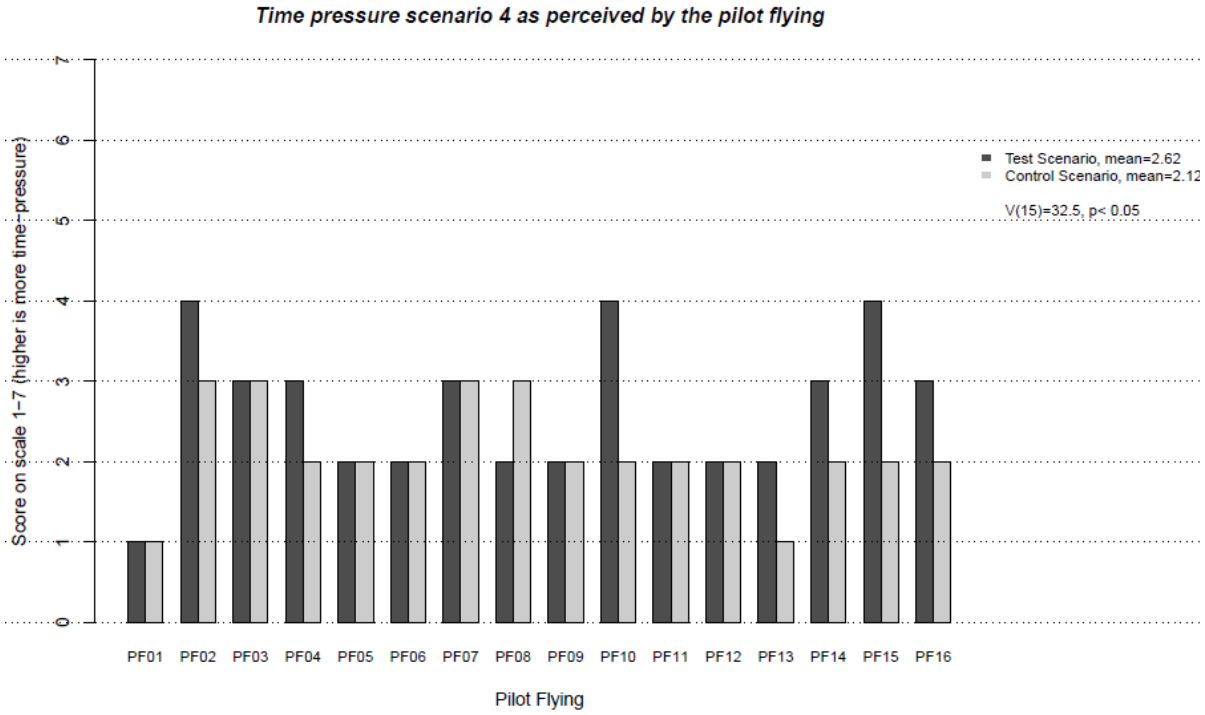

Figure 36: The time pressure in scenario 4 as perceived by the pilot flying.

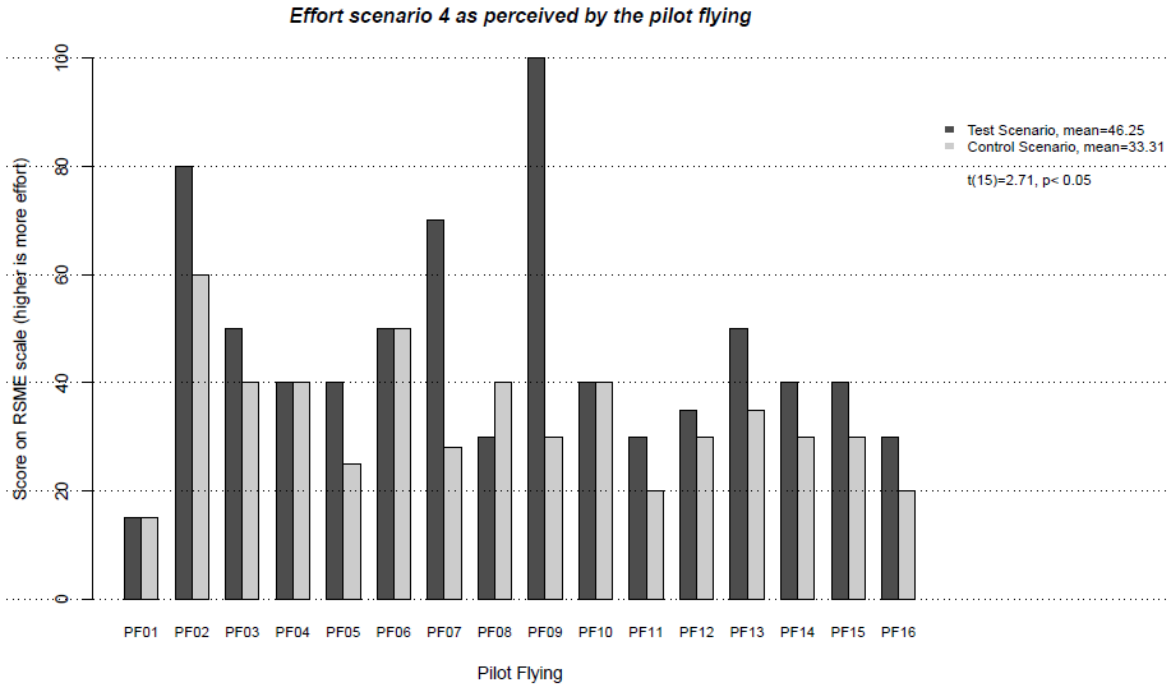

Figure 37: The effort as perceived by the pilot flying in scenario 4.

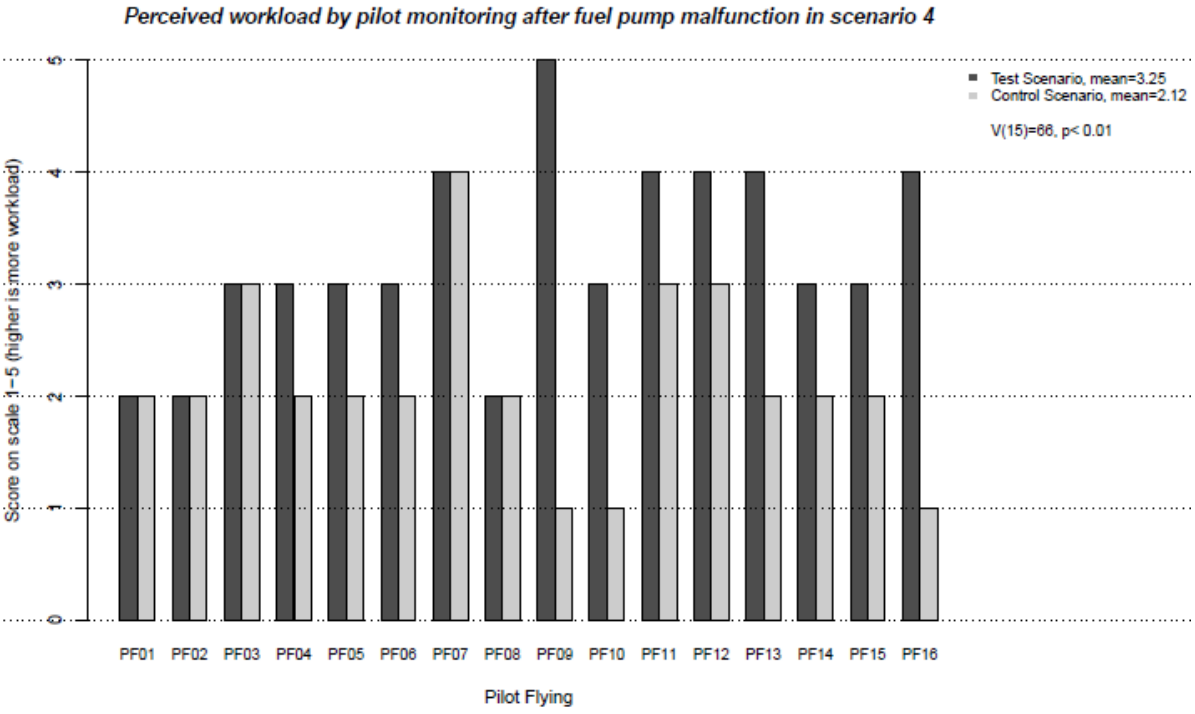

Figure 38: The workload of the pilot flying as perceived by the pilot monitoring after the fuel pump malfunction in scenario 4.

7.3.2 Conclusion Hypothesis 32

The hypothesis that the test scenarios were experienced by the pilot flying as costing more effort was supported for scenarios 1 and 4. The hypothesis on the pilots experiencing the test scenarios to have more tasks was supported by scenarios 1, 3 and 4. The hypothesis about more time pressure for the test scenarios as experienced by the pilot flying was supported for scenarios 1 and 4.
The pilot monitoring rated the task load of the pilot flying higher in scenarios 1, 3 and 4.
This shows that, except for scenario 2, the test scenarios lead to a higher task load as they were designed to do. However, for scenario 3, this higher task load is not consistently experienced as being higher. The pilot flying's rating of the effort as being higher in the test scenario in scenario 3 was not significantly different to the rating in the control scenario. For that reason, scenario 3 will not be taken into account in the following analyses.

Scn	Nr. of tasks	Time pressure	Effort	Task load
1	Significant (p<0.001)	Significant (p<0.001)	Significant (p<0.001)	Significant (p<0.05)
2	Not sign. (p=0.97)	Not sign. (p=0.12)	Not sign. (p=0.26)	Not sign. (p=0.12)
3	Significant (p<0.05)	Not sign. (p=0.18)	Not sign. (p=0.65)	Significant (p<0.01)
4	Significant (p<0.05)	Significant (p<0.05)	Significant (p<0.01)	Significant (p<0.01)

Table 5: Overview over the difference of the subjective experience of the independent variables between the test and the control scenario.

In Table 5, you find an overview of the difference of the subjective experience of the independent variables between the test and the control scenario.

Only in test scenario 1 the objective measures for tasks and time pressure were rated as being closer to high (7) than to low (1), 4.4 and 4.25 respectively. Also the pilot flying rated the effort of test scenario 1 between *rather much effort* and *considerable effort* while the other scenarios were all rated between some effort and rather much effort. Only the task load as rated by the pilot monitoring was high for test scenarios 1, 3 and 4, though highest for scenario 1. Following this, the only scenario we expect to see cognitive lockup is scenario 1, because the experienced task load and time pressure should both be high before cognitively lockup exhibits.
For scenario 2, we do not expect cognitive lockup, as there is no significant difference in the rating of the variables between test and control scenario.

Even though scenario 4 has a significant difference in rating, the value of the rating was too low; no high effort has been experienced by the pilot flying. As a consequence, also in scenario 4, no cognitive lockup is expected.

7.4 Analysis of H17 and H36: Objective data analysis Cognitive Lockup

From the subjective data, it could be derived which variable manipulation has been successful. For an overview, see Table 5. For the conclusion, please see section 7.3.2.

In this section, we will evaluate hypotheses 17 and 36, which are:

- H17 : Reaction time of pilots to visual events (AHMI popup box) depends on workload
- H36 : The pilots will show Cognitive Lockup at moments in the scenario when there are multiple tasks with similar priorities when the pilot is executing a task with a high mental workload

In D4.6, we divided hypothesis 36 in two hypotheses. Cognitive lockup was defined as 'strong' and 'weak' cognitive lockup, with 'strong' meaning that a task switch is actually done too late (and thus not corresponding to normative behaviour anymore), and 'weak' cognitive lockup meaning that the task switch is done significantly later than when no cognitive lockup occurs.
The two hypotheses are:

1. If the TSS is "high" and the TOC is "high" the switch to another (higher priority) task is not done on time.
2. If the TSS is "high" and the TOC is "high" the experienced effort is "high" then the reaction time to switch to another (higher priority) task than in situations where TSS and TOC are lower.

In the following, we will first have a look at the second hypothesis, and will evaluate the reaction time of the pilots. Only if the reaction time differs significantly, cognitive lockup occurred and the first hypothesis will be relevant to evaluate.

For evaluating hypotheses 17 and 36, we need the scenarios to differ significantly in the dimension *workload*, which in turn is built up by the two variables *number of tasks* and *time pressure.* This has been evaluated with hypothesis 32, in section 7.3. As a conclusion, we expect cognitive lockup in the test scenario of scenario 1, no cognitive lockup in the test scenario of scenario 2. For scenario 4, there was a significant difference in the experience of effort, however, this effort was not rated very high in the test scenario. For that reason, we expect a difference in reaction time, although not a significant one. We expect scenario 3 to not have a clear result, as the manipulation of the variables has not been experienced in a consistent way.

7.4.1 Simulator data

For the objective simulator data, the reaction time between uplink and looking at AHMI, the reaction time from looking at AHMI to action and the complete duration were measured. Unfortunately, the eye gaze data was not stable enough to use as a measure, probably due to the dark environment in which the experiment took place. The data of pilots 3 and 4 was missing for scenario 1. For scenario 2 the data of pilots 3 and 16 was missing. The data of pilots 3,4,5 and 6 were missing for scenario 4.

We also looked at the pupil dilation of the pilots because it is known that a dilated pupil can be a task load indicator [4]. Unfortunately this data could not be used; the data was not complete enough. The pilots were in a dark environment causing their pupils to be dilated the whole time. In Figure 39, an example of the measurement is provided, showing the pupil diameter of pilot 6 in scenario 1.

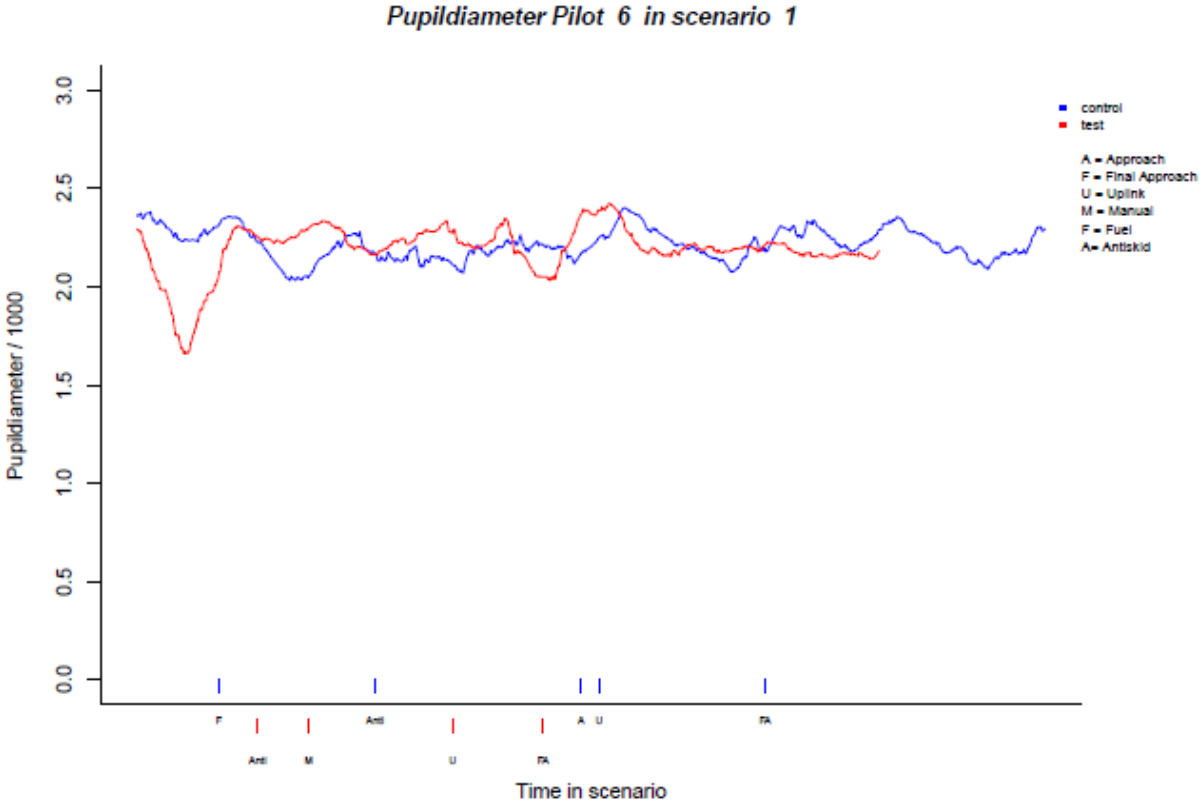

Figure 39: The pupil diameter of pilot 6 in scenario1.

7.4.1.1 Scenario 1

For scenario 1, as reaction time, we measured the complete time from time from uplink to sending the implemented changes to the ATC (Figure 40). The difference in reaction time between test and control scenario was significant ($p < 0.001$).

7.4.1.2 Scenario 2

For scenario 2 we measured the complete time from time from uplink to sending the implemented changes to the ATC (Figure 41). There was no significant difference between the test and control scenario, $p = 0.95$.

7.4.1.3 Scenario 3

For scenario 3 the error that we expected the pilot to make in case of cognitive lockup was monitoring too late or not at all of the altitude constraints.
As this involves no specific eye gaze patterns or button presses, it can only be deduced from video data. We did not do this for this deliverable. As specified in Table 5, the manipulation of the variables was not successful in scenario 3, and we did not succeed in having a consistent variation in the variables (so not even no variation). For this reason, the analysis of scenario 3 for cognitive lockup has been decided to be not applicable.

7.4.1.4 Scenario 4

For scenario 4 the time from uplink about fuel pump malfunction to the time that the pilot changes the altitude is taken (Figure 42). The difference between test and control scenario is not significant ($p=0.52$).

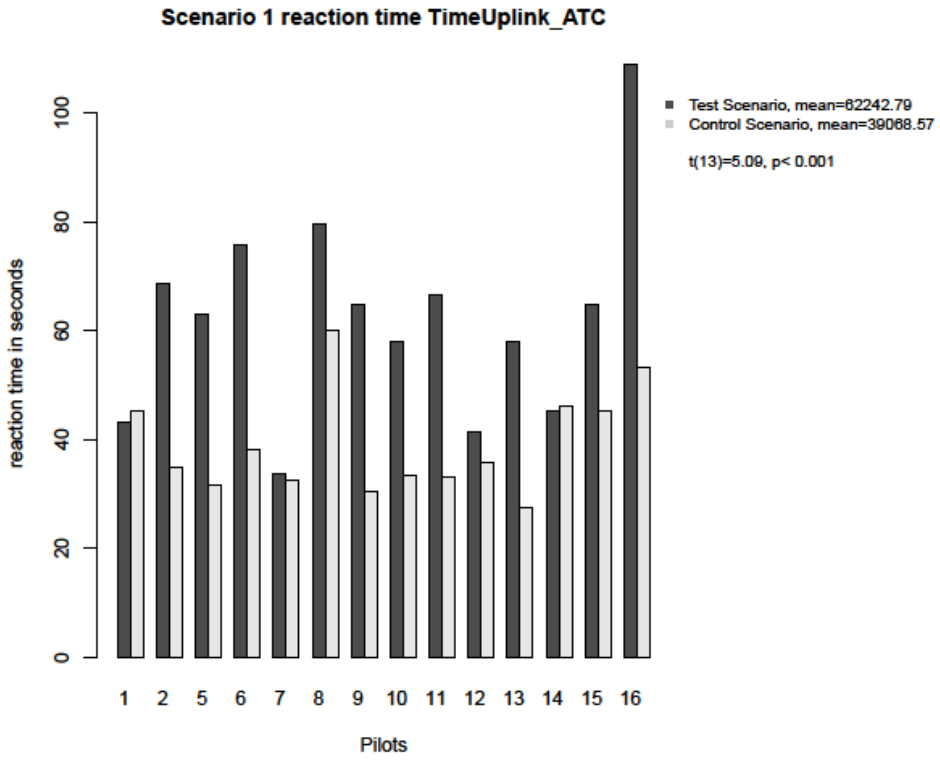

Figure 40: Reaction time between the uplink is received to sending the implemented changes to the ATC in scenario 1.

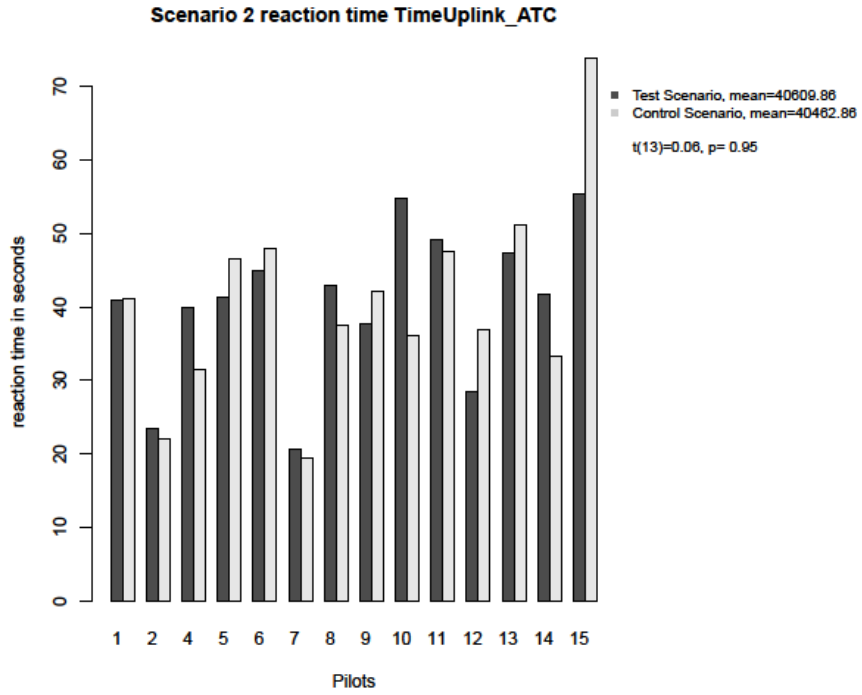

Figure 41: Reaction time between the uplink is received to sending the trajectory to ATC in scenario 2.

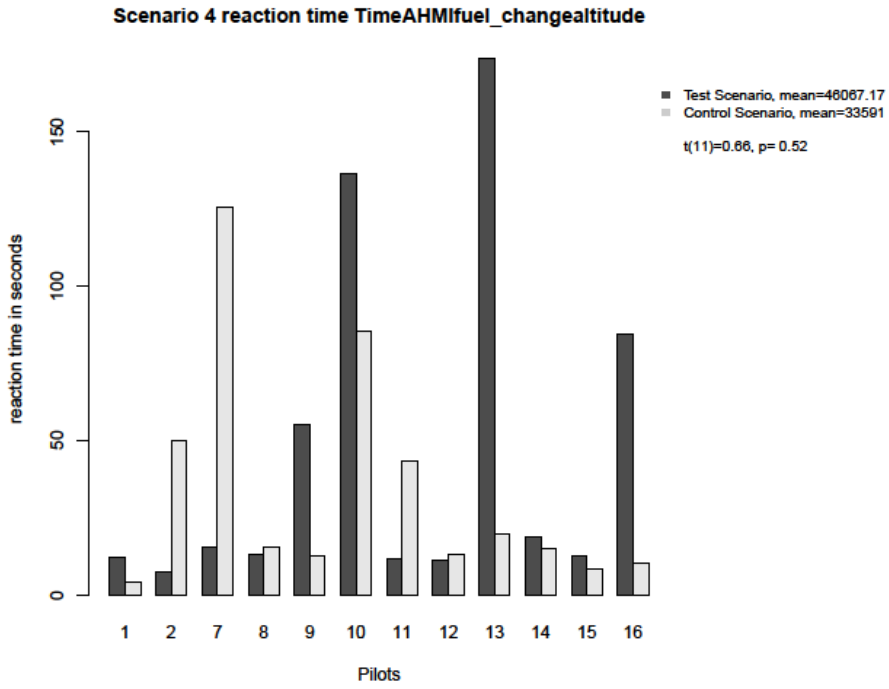

Figure 42: Reaction time between the uplink to the time the pilot changes the altitude in scenario 4.

7.4.2 Model data and comparison with simulator data

Due to some problems with running the integrated model in a batch mode, we do not have very many runs of the model of the scenarios at this moment.
As described above (see section 7.3.2), we identified scenario 1 to be the most interesting scenario for the evaluation of the model. The main reason for this is that the manipulation of the variables has been successful and the pilots actually experienced a significant difference in workload. For that reason, in the following, we concentrate on the analysis of scenario 1.

With the model, there are four runs for scenario 1 for the test scenario and one for the control scenario.
For test scenario 2, we have 4 runs and none for the control scenario. In scenario 2, we do not expect any cognitive lockup. The main reason for the runs is to evaluate whether the reaction time corresponds to the reaction time of the pilots.

In scenario 1, the reaction time between the time the uplink is received to when the pilot negotiates a new trajectory is evaluated. For the overview of the reaction times

for the model in scenario 1, see Figure 43. The mean reaction time of the model for the uplink in scenario 1 in the test condition was 31.7s and the reaction time for the control scenario was 15.2s. The reaction times of the real pilot data are respectively 62.2 and 39.1s. The reaction time of the model thus seems too fast, although still a clear difference can be seen between the two conditions.
If we only take the test condition into account, the model should be twice as slow. We cannot really take the control scenario into account, because we cannot establish a mean on basis of one run.

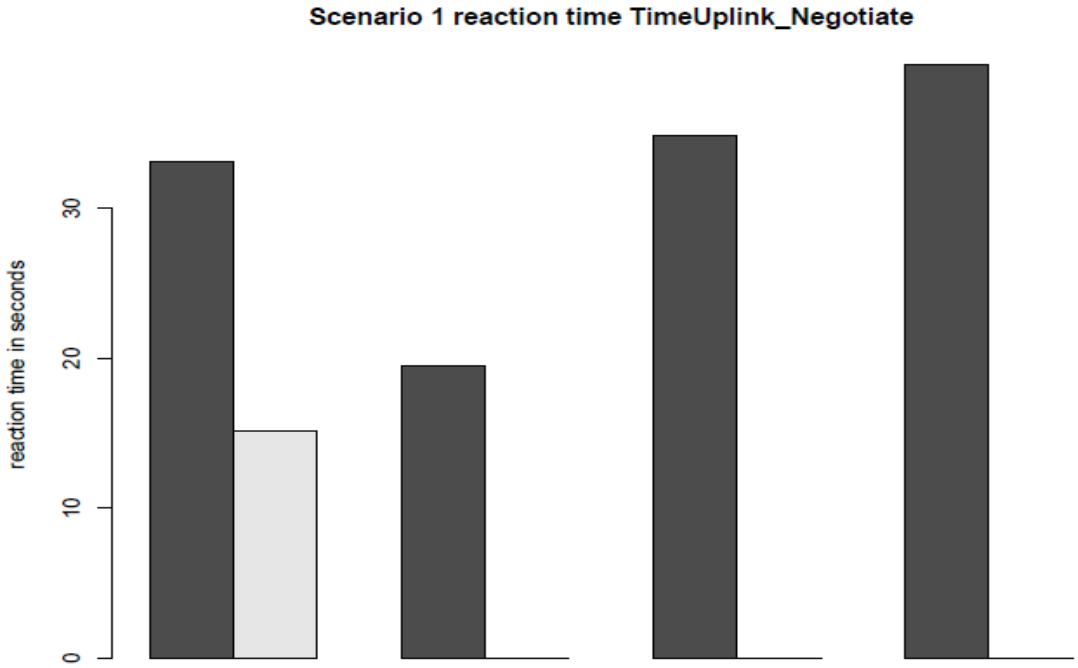

Figure 43: Reaction time between the time the uplink is received to when the pilot negotiates a new trajectory in scenario 1.

In scenario 2, the reaction time of the model between the time an uplink is received to the time the pilot starts negotiating a new trajectory is evaluated. For the overview of the reaction times for the model in scenario 2, please see Figure 44.
The mean reaction time for scenario 2 in the test condition was 28.8s while this was 40.6s for the pilot data. So also for scenario 2 the model is too fast, although the variability of the reaction times corresponds quite well with the variability of the reaction time of the real pilots.

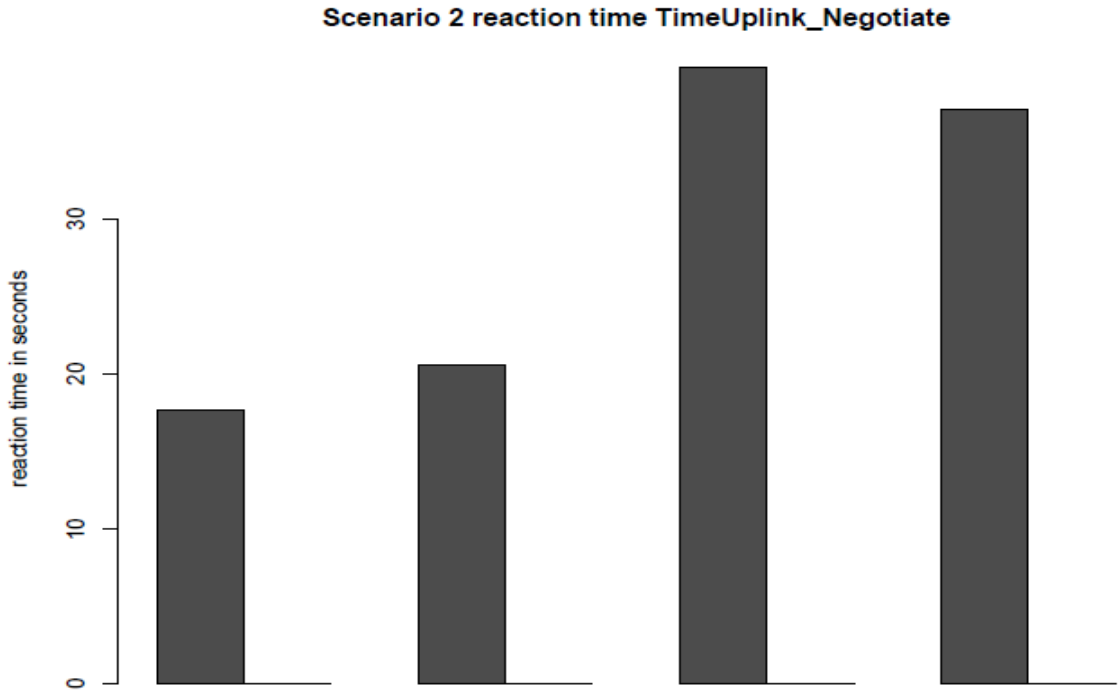

Figure 44: Reaction time between the time an uplink is received to the time the pilot starts negotiating a new trajectory in scenario 2.

7.4.3 Conclusion

In this section, we evaluated hypotheses 17 and 36, which are:

- H17 : Reaction time of pilots to visual events (AHMI popup box) depends on workload
- H36 : The pilots will show Cognitive Lockup at moments in the scenario when there are multiple tasks with similar priorities when the pilot is executing a task with a high mental workload

As described in the introduction of section 7.4, we have two different definitions of cognitive lockup; one of them, 'weak' cognitive lockup, is about the reaction time (see D4.6), whereas 'strong' cognitive lockup is about actually deviating from normative behaviour. The 'weak' definition thus lies close together with H17 (only that it is about a *longer* reaction time in case of cognitive lockup). As all scenarios have multiple tasks with similar priorities (see D4.6), the focus lay on the workload. We have evaluated whether 'weak' cognitive lockup occurred. For the 'strong' analysis, the data cannot be taken directly from the simulation data, but the videos need to be watched to determine whether actual errors have been made. This has not been done for this deliverable, but we have focused on 'weak' cognitive lockup.

Scenario 1

The results show that the reaction time of the pilots between noticing the uplink and finishing the action in test scenario 1 is significantly higher than in control scenario 1. This corresponds to our expectations (see section 7.3.1). In addition, the number of tasks and time pressure was also experienced to be high by the pilots.
The model simulates the variability in reaction time between the pilots quite well. In addition, there is a difference in reaction time between the test scenarios and the control scenario, just as in the pilot experiments. However in general, the model reacts too fast.

Scenario 2

The results also show that for scenario 2 where the experienced task load (effort, time pressure and number of tasks) was equal for the test and control scenario the reaction time was also almost equal ($p=0.97$).
The model simulates the variability in reaction time between the pilots quite well. The comparison with the model data shows that the model is too fast.

Scenario 3

Scenario 3 was decided to be not applicable, as the manipulation of the variables was not successful.

Scenario 4

For scenario 4 for instance the subjective measures showed a significant difference between test and control, but even so the experienced task load was not high. This resulted in a reaction time that was longer, but not significantly longer between the test and control condition.

In general, the results show that it is difficult to create an experiment that creates enough workload to induce cognitive lockup. On the other hand it also shows that the subjective measures do indicate the change on cognitive lockup.

8 New Requirements for the Cognitive Model

In this section, the requirements are described that can be identified after comparing the results from the pilot experiments with the results from the model experiments.
We first describe requirements that come forth from the analysis of the hypotheses about the basic capabilities, followed by the requirements that can be derived from the analysis of the hypotheses about the error production mechanisms.

8.1 New Requirements from the Analysis of the Basic Capabilities

After analysis of basic capabilities we have identified possible improvements for the pilot model. The analysis of improvements made after cycle 1 with regard to visual behaviour show that no further improvements are needed. With regard to Task Execution Time on the AHMI improvements are needed. Analysis has shown that the model performs too fast. Slowing down the model can be made by modifications of the basic skills regarding Fitt's Law. These skills should be tuned with regard to a trackball device for operations on the AHMI. These modifications will improve the degree of realism are relevant for the scenarios, because longer execution times will affect the attention that the model can spend to other important tasks. This improvement will have low implementation effort.

8.2 New Requirements from the Analysis of the Error Production mechanisms

8.2.1 Requirements for Learned Carelessness

The implementation of the parameter effort showed significant reduction of false predictions of LC, and seems to be a reasonable addition to the parameter frequency of events. The data showed that their risk is not a suitable parameter, but this needs further assessment.
Requirements:

1) Integration of the ISAAC rule learning mechanism with the new memory learning mechanism of HUMAN: When the memory learning mechanism has evolved for certain conditions, the rule could be simplified. The simplified rule then is undergoing the rule selection process, and a learning process of rule strength, which is dependent on success and failures.
2) Improvement for the visualisation of LC results. The memory picture is hard to interpret, as many associations could be evolve. Depicting such heavy memory structures is not easy. An improvement could be to highlight the conditions that have been identified as prone to LC. This could be either done in CASCaS, or in PED.

8.2.2 Requirements for Selective Attention

The current selective attention process is based on probabilistic choices. The probabilistic choice is dependent on the neighbourhood of the instrument where the event (motion of flashing) takes place. This mechanism should be improved in the following ways:

1) Dynamic calculation of the saliency of the neighbourhood. In the current implementation the information if an instrument is statically described in the topology. In fact, the dynamicity of an instrument changes over time, e.g. a PFD is colourful, but does not always show motion (e.g. when flying straight

ahead, the PFD is static). Therefore, this information should be interfered online from the current status of the instrument.
2) Adaption of the probabilistic choice, based on the online value of the dynamicity of the neighbourhood.

8.2.3 Requirements for Cognitive Lockup

The speed of the model should be decreased to be more according to the pilot data. For this it would be good to look at different sections of the pilot data and compare them to the same sections in the model data. This would make the decrease in speed be based on actually results. By doing this the results will also be more generic.

Further are several improvements possible on the mechanism of cognitive lockup itself:

- At the moment, we hypothesized (and set up the calculation of the workload in the model accordingly), that the workload depends on the two parameters *time pressure* and *number of tasks* (for the motivation, please see D4.6). The results of the evaluation of H32 support this thesis. However, at the moment, in the model, these two variables have the same influence on the workload. This does not need to be the case, and needs to be refined. More model runs are needed with a variation of the values for the parameters, and more experimental research is needed to determine on the importance of the two variables for workload.
- The variable *time pressure* is calculated at the moment according to a normative knowledge on how long there is in general for a task. It is not calculated dynamically according to the actual scenario situation. This needs to be improved.

9 References

1. Mumaw, R. J. , Sarter, N. B., Wickens, C. D.; Analysis of Pilot's Monitoring and Performance on an Automated Flight Deck; Presented at the 11th International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University. 2001.
2. Neerincx, M.A. Cognitive task load design: model, methods, and examples. In: E. Hollnagel (eds.), *Handbook of Cognitive Task Design*. Chapter 13 (pp. 283-305). Mahwah, NJ: Lawrence Erlbaum Associates. 2003.
3. Zijlstra, F.R.H. *Efficiency in Work Behaviour: a Design Approach for Modern Tools*. Delft: Delft University Press. ISBN 90-6275-918-1. 1993.
4. Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., and Jung, T.-P. Task Performance and Eye Activity: Predicting Behavior Relating to Cognitive

Workload. Aviation, Space, and Environmental Medicine; 78(5, Suppl.): B176-85. 2007.