



Executive summary

A coherent impression of the pilot's situation awareness

Studying relevant human factors tools

Problem area

A flight simulator experiment was set up to study relevant Human Factors tools for Situation Awareness assessment of pilots.

Description of work

A specific scenario was designed in which a malfunction of the aircraft was introduced during flight and slowly progressed: an Indicated Air Speed discrepancy. Pilot behaviour was studied using eye tracking and subjective, self-report rating scales while pilots tried to figure out the correct air speed.

Results and conclusions

As it turned out, these Human Factors tools together provided a coherent impression of the pilots' compromised Situation Awareness, covering all three levels of Endsley's Situation Awareness definition.

Applicability

Individual eye movement metrics alone provide an insufficient picture of operator Situation Awareness, but when purposefully combined with subjective, self-rating metrics, offer a more comprehensive look at operator Situation Awareness.

Report no.

NLR-TP-2009-627

Author(s)

H. van Dijk
G.K. van de Merwe
G.D.R. Zon

Report classification

UNCLASSIFIED

Date

May 2010

Knowledge area(s)

Training, Simulatie en Operator Performance

Descriptor(s)

Human factors
Situational awareness
HILAS

A coherent impression of the pilot's situation awareness
Studying relevant human factors tools



NLR-TP-2009-627

A coherent impression of the pilot's situation awareness

Studying relevant human factors tools

H. van Dijk, G.K. van de Merwe and G.D.R. Zon

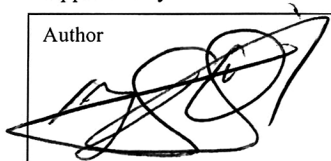
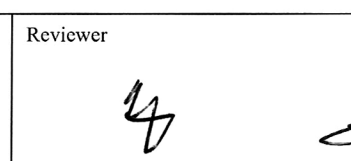
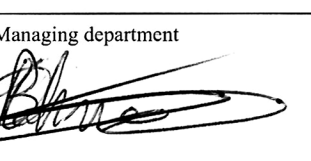
This report is based on an article to be published in the Journal of Aviation Psychology.

The contents of this report may be cited on condition that full credit is given to NLR and the authors.

This publication has been refereed by the Advisory Committee AIR TRANSPORT.

| | |
|-------------------------|-----------------------------------|
| Customer | European Commission |
| Contract number | ---- |
| Owner | National Aerospace Laboratory NLR |
| Division NLR | Air Transport |
| Distribution | Unlimited |
| Classification of title | Unclassified |
| | May 2010 |

Approved by:

| | | |
|---|--|--|
| Author  | Reviewer  | Managing department  |
|---|--|--|

Summary

A flight simulator experiment was set up to study relevant Human Factors tools for Situation Awareness assessment of pilots. A specific scenario was designed in which a malfunction of the aircraft was introduced during flight: an Indicated Air Speed discrepancy. Pilot behaviour was studied while pilots tried to figure out the correct speed. Eye movement metrics alone provided an insufficient picture of pilot Situation Awareness, but when purposefully combined with subjective, self-rating metrics, they offered a more comprehensive look at Situation Awareness, covering all three levels of Endsley's Situation Awareness definition.

Contents

| | | |
|----------|--------------------------------|-----------|
| 1 | Introduction | 4 |
| 1.1 | Situation awareness assessment | 4 |
| 1.2 | Flight simulator experiment | 5 |
| 2 | Methods | 7 |
| 2.1 | Participating pilots | 7 |
| 2.2 | Flight simulator | 7 |
| 2.3 | Flight scenario | 8 |
| 2.4 | SA measurement techniques | 8 |
| 2.5 | Data analysis | 9 |
| 3 | Results | 10 |
| 4 | Discussion | 13 |
| | References | 16 |

1 Introduction

Over the past two or three decades, the concept of Situation Awareness (SA) has received considerable attention from the Human Factors (HF) research community. Although, originally a term used within (military) aviation, SA has developed into a major concern in many other domains where people operate complex, dynamic systems (e.g. maintenance, air traffic control, medical systems, and nuclear power industry). Achieving SA is one of the most challenging aspects of these operators' jobs and is central to good decision making and performance. This stems largely from a growing concern with the effects of widespread automation and advanced information systems on the ability of humans to take in and comprehend exactly what is going on without becoming confused, overloaded or error-prone. As a consequence, valid and meaningful measures of SA are required to help us assess the design and use of complex systems in simulations and operational settings. The current article describes a flight simulator experiment that was set up to study relevant HF tools for SA assessment of pilots.

1.1 Situation awareness assessment

SA defined. Many theories of SA have been developed. The most commonly used and widely cited theory of SA is the three-level model of SA proposed by Endsley (1995a). Endsley defines SA as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1995a, p. 36). Inherent in this definition is the operator's notion of what is important. SA is frequently defined in operational terms. For a given operator, therefore, SA is defined in terms of the goals and decision tasks for that job. The pilot does not need to know everything (e.g. the co-pilot's shoe size), but does need to know a great deal of information related to the goal of safely flying the aircraft. The “elements” of SA vary widely between domains, however, the nature of SA and the mechanisms used for achieving SA can be described generically.

The three-level model of SA divides SA into three hierarchical activity levels (for more detailed information on the three-level model of SA, see Endsley, 1995a; Stanton et al., 2006; Vidulich, 2003). For instance, in an aircraft environment, pilots must be aware of critical flight parameters, the state of their on-board systems, their own location and the location of important reference points and terrain, and the location of other aircraft. This information forms the elements they need to perceive for good Level 1 SA (i.e. “perception of the elements in the environment”). But a great deal has to do with how the operators interpret the information they take in. Pilots need to comprehend that the displayed altitude is below their assigned altitude. This understanding forms their Level 2 SA (i.e. “comprehension of the elements and their meaning”). At the highest level, Level 3 SA (i.e. “projection of future status”), their

understanding of the state of the system and its dynamics can allow them to be able to predict the near future. Pilots may predict that in case of a diversion, due to bad weather, they may experience fuel shortage and, therefore, need to look out for an alternate airport.

SA measured. The assessment of SA is often used throughout the design lifecycle, either to evaluate the effect of (novel) technologies and training interventions upon SA, to assess SA in existing operational systems, or to examine factors that affect SA. Measurement is needed for systematic improvement of human performance, either by training or by systems design. Several HF tools have been established for the measurement of SA (for a complete overview of SA measures, see Gawron, 2008; Stanton et al., 2006). Taylor et al. (1995) suggested approaches that include performance-based metrics, physiological indices, memory probe measures of SA knowledge and subjective, self-report ratings. In a review of SA measurement techniques, Endsley (1995b) described the same approaches more or less, including physiological measurement techniques, performance measures, external and imbedded task measures, subjective rating techniques (self- and observer-ratings), questionnaires (post-trial and on-line) and the so-called freeze technique.

The construct of SA is complex and comprises many aspects. As such SA should not be assessed using a single tool. Here, the strategy entails the use of different HF tools and see if the different results converge into the same direction (also referred to as the converging evidence principle). For instance, ask pilots to rate their own SA and compare these ratings with the ratings from a subject matter expert who monitored the flight on video. These tools together should provide a coherent impression of the SA in this particular setting. Furthermore, the choice for a certain SA measurement technique should be dependent on the level of SA (according to Endsley, 1995a) that is being assessed. For instance, Endsley's definition of SA claims that pilots first have to notice a particular phenomenon in order to become aware, understand and project into the future. This first step (i.e. "perception of elements in the environment") can be measured by eye tracking. The eye tracker illustrates where the pilots focus their attention on. As such the eye tracker could be a helpful tool to measure Level 1 SA.

1.2 Flight simulator experiment

The primary goal of the current flight simulator experiment was to study relevant HF tools for SA assessment of pilots and to determine if these tools together provide a coherent impression of the pilots' SA. A specific scenario was designed in which a malfunction of the aircraft was simulated: an Indicated Air Speed (IAS) discrepancy. The malfunction was introduced unbeknownst to the pilots during flight and slowly progressed over time while researchers monitored if and how pilots detected the IAS discrepancy and figured out the correct air speed.

Pilot behaviour was studied during the scenario using eye tracking and two different subjective, self-report rating techniques (on-line and post-trial ratings). It was expected that the different HF tools together would provide a more coherent impression of the pilots' SA than the individual measures separately.

Eye tracking. In many studies, the direction of an operator's viewing direction has been monitored to determine what instruments, displays or information sources were used. The assumption should be made here that the operator attends to the information his eyes are directed toward; i.e. eye movements are an indicator of visual attention (for a review, see Rayner, 1998). Researchers at NASA-Langley Research Center have conducted numerous studies in which eye tracking was used in a general aviation simulator (summarized in Harris et al., 1986). They proposed two measurement techniques for display evaluation: (1) comparing histogram plots of dwell times for different instruments and (2) measuring scan pattern changes as the difficulty of a task increases. The two analytic techniques provided converging evidence that one instrument was superior to the other.

Part of the current research focused on the question if eye tracking is a useful HF tool to understand how SA is maintained / obtained. In the current research, where pilots are trying to detect a simulated malfunction, the dwell times for different instruments might reveal where the pilot's primary focus of attention is. This typically reflects a Level 1 SA: perception of the IAS discrepancy. It was expected that pilots would spend more time looking at instruments where information about the malfunction may be found, at the cost of looking at other instruments where they would have looked when there was no malfunction. Additionally, the randomness of the pilots' scanning pattern might reveal the search strategies of the pilots, trying to detect and interpret the malfunction. This probably tells us more about the Level 2 SA: comprehension of the IAS discrepancy. It was expected that the scanning pattern would be more random after the pilots discovered the malfunction because they were searching for the solution.

Subjective ratings. The simple rating technique called Instantaneous Self Assessment (ISA) rating scale (Castle & Legatt, 2002) was used in the current research. The ISA rating scale was originally developed to evaluate mental workload of air traffic controllers and pilots. However, instead of evaluating workload, the current research used the ISA technique to measure the pilots' overview of the situation on that particular moment in the scenario. This comprised several on-line ratings of the pilots. The ISA ratings do not specifically measure a certain level of SA; it generally focuses on the idea of "what is going on". The research question concerning the ISA rating scale was if ISA ratings provide generic insight in the course of the pilots' SA during the scenario. Knowing if and when SA degenerates and/or improves can be of

great value to the interpretation of other measures. ISA ratings were expected to degenerate when pilots became aware of the simulated malfunction.

The Crew Awareness Rating Scale (CARS; McGuinness & Foy, 2000) is a SA assessment technique (this comprised a post-trial rating scale) that is based upon the three-level model of SA (Endsley, 1995a). The scale is designed to elicit ratings based upon ease of identification, understanding and projection of task SA elements (i.e. Levels 1, 2 and 3 SA). Stanton et al. (2006) claim that CARS could be used in conjunction with on-line techniques – such as the ISA rating scale – to ensure comprehensiveness. Using CARS only an overall SA rating is acquired at the end of the task, rather than several ratings during the course of the task. This could potentially inhibit the usefulness of the output, not knowing the course of the pilots' SA during the scenario, and not being able to pinpoint specific problems. The research question concerning CARS was if this is a useful SA assessment technique, in conjunction with an on-line subjective rating scale, being the ISA rating scale. The overall CARS rating was expected to be relatively low due to the hampered SA in the scenario.

2 Methods

The current experiment was part of a larger study on the applicability of Human Factors tools in the design, evaluation and operation of aviation systems (for an overview of this study, see Zon & Van Dijk, 2009).

2.1 Participating pilots

Six crews of each two airline pilots (i.e. a captain and a first officer) participated in the current experiment: one Italian, two Spanish and three Dutch crews. The average age of the participating pilots was 38 years (standard deviation SD 6.0 years). All pilots were active and qualified to fly an Airbus A320. On average they had 7450 hours of flight experience (SD 4850 hours).

There were two pilot roles in the simulation: pilot flying (PF), carried out by the first officer, and pilot not flying (PNF), carried out by the captain. The pilot tasks match normal operations.

2.2 Flight simulator

GRACE (Generic Research Aircraft Cockpit Environment) is a generic flight simulator, representing a modern large two-engine fly-by-wire airliner (Heesbeen et al., 2006). GRACE has a number of standard configurations. For the current experiment the Airbus A320 configuration was selected. A high fidelity simulator such as GRACE allows researchers to

perform realistic experiments in a fully controlled environment. Before the experiment started, extensive practice runs were conducted by the pilots to familiarise with the GRACE simulator.

2.3 Flight scenario

All crews performed one run of the same simulated flight. The simulated flight consisted of a trip from London Heathrow to Amsterdam Schiphol (starting in cruise, ending after touchdown). A specific scenario was designed simulating a certain malfunction of the aircraft during the flight: an IAS discrepancy was introduced. The discrepancy was indicated by the two available Primary Flight Displays (PFDs); i.e. one display showed the correct air speed while the other showed a lower, false air speed (see Figure 1 for a PFD with the speed tape circled in red). Once the discrepancy was initiated by the simulator after about 10 minutes in flight, it slowly progressed over time while researchers monitored if and how the pilots detected the discrepancy, and if and how the pilots figured out what the correct air speed was. As far as the crew was concerned, they were flying a normal flight until the malfunction was detected. The flight duration of this scenario was 25 minutes.

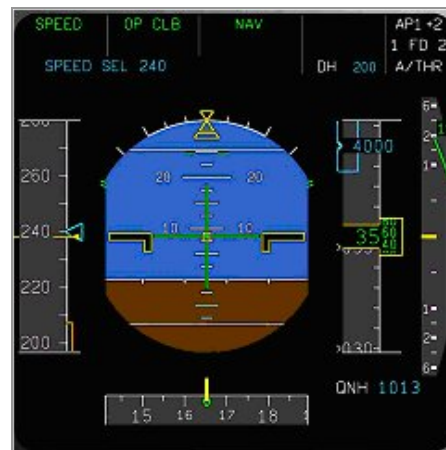


Figure 1. PFD with the speed tape circled on the left vertical bar

In addition to the current flight scenario, where the IAS discrepancy was introduced, a reference scenario was flown by all crews. This reference scenario was similar to the first part of the IAS discrepancy scenario; i.e. before the malfunction was initiated. The duration of this reference scenario was 10 minutes. The sequence in which both scenarios were flown was randomized.

2.4 SA measurement techniques

Pilot behaviour was studied using the following HF tools:

Eye tracker. Eye tracking data were collected using the Applied Science Laboratories 6000 eye tracker, together with an Ascension Technologies optical head tracker. Both pilots wore a headband on which the optonics were mounted. During flight, all eye activity was recorded. To facilitate data analysis of the current flight scenario, five relevant Areas-of-Interest (AoIs) were defined for the PF and PNF: Primary Flight Display (PFD), cross check (i.e. other PFD), Navigation Display (ND), Electronic Centralized Aircraft Monitoring (ECAM) system, and “other” (e.g. outside).

Dwell time and entropy were used as eye tracking measures for the PF and PNF. Dwell time provides information regarding the amount of time spent viewing the different AoIs. Entropy is a measure of randomness of the viewing pattern (for the used calculation method, see Ellis & Stark, 1986).

ISA. The five-point ISA rating scale was used in the current research to measure the pilots’ overview of the situation on that particular moment in the scenario; i.e. the ratings were self assessed during the flight (on-line ratings). The PF and PNF were asked (every 2 minutes) to respond to the rating scale presented on the touch screen display by assessing his/her (individual) current situation overview (5 being very high and 1 being very low).

CARS. The CARS was administered post-trial to assess SA. The CARS identifies four common components of SA: (1) perception: detecting and recognising current facts and data; (2) comprehension: making appropriate interpretations of the facts and data; (3) projection: making realistic predictions of future developments; and (4) appreciation: appreciating implications for goals, decisions and actions. For each aspect of SA there are two rating scales, one addressing the perceived accuracy of the awareness and the other addressing the perceived difficulty of the processing involved, which is an aspect of workload. To enable this assessment, the questions used with the scale were framed so as to prompt the pilots to consider whether or not they think they have “good” SA or are finding it “easy to maintain” SA. Each question came with a rating scale, plus a “not applicable” response option.

2.5 Data analysis

It was expected –as the IAS discrepancy progressed over time– that most of the crews detected the malfunction earlier than after the 2 minutes on which the aircraft would automatically send out a warning. As it turned out, none of the six crews discovered the specific discrepancy on both PFDs before an auditory warning was heard and the ECAM system indicated the air speed discrepancy. Consequently, the analysis focussed only on the time that the crew needed to figure out the correct air speed, and not on the time they needed to detect the IAS discrepancy (i.e. the ECAM system already indicated the discrepancy). The period after the warning until the

moment of determination of the correct air speed (as observed by the experiment leader) is referred to as post-period. In the analysis of the eye tracking measures (dwell time and entropy) and ISA ratings, this post-period is compared to a pre-period that has the same length as the post-period and takes place immediately before the onset of the IAS discrepancy (see Figure 2 for an illustration of this time-line).

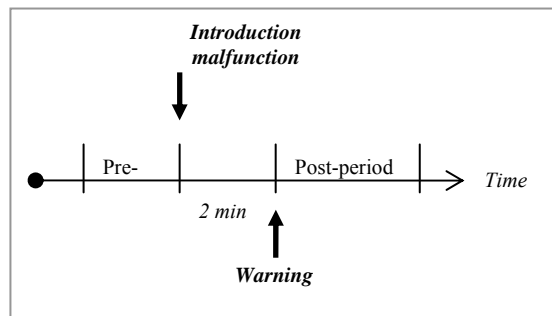


Figure 2. Illustration of the time-line for the analysis

Concerning the analysis of the CARS ratings, comparisons are made between post-trial ratings for the reference scenario and the IAS discrepancy scenario.

3 Results

An α of 5% was used for significance testing and Cohen's d was used as a measure of effect size.

Dwell times. The dwell times (in percentage) on the different AoIs (with exception of the AoI "other") in the pre- and post-period are shown in Figure 3.

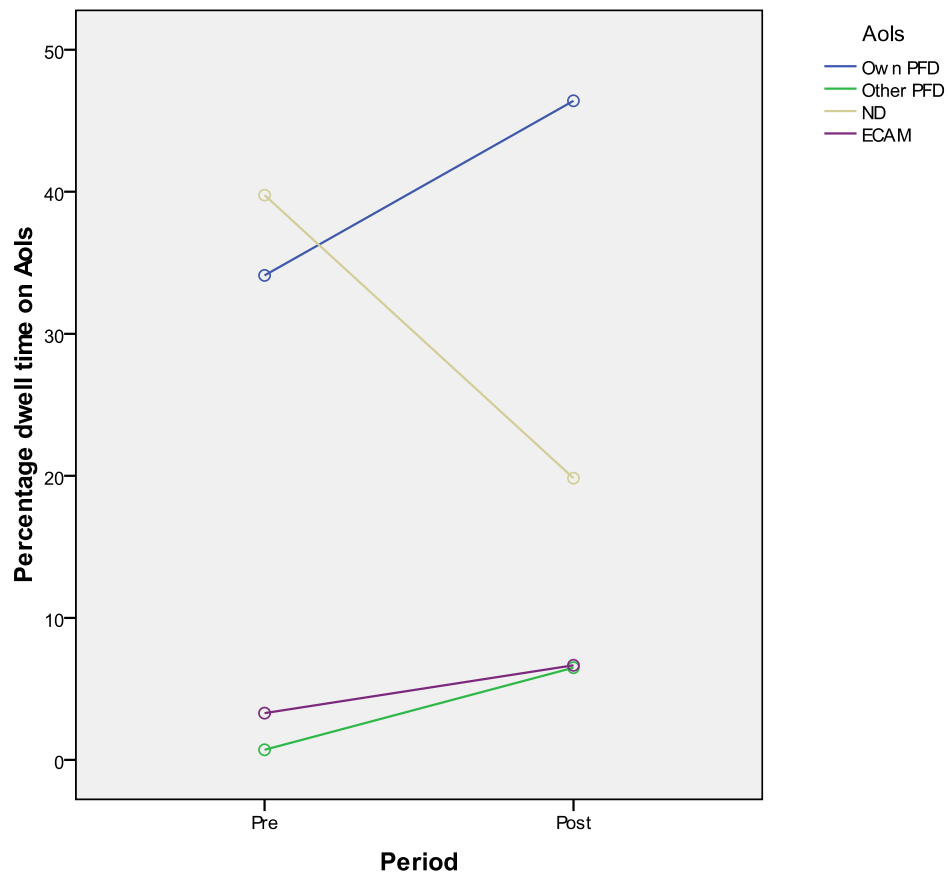


Figure 3. Percentage dwell time on pilot's own and other PFD, ND, and ECAM in the pre- and post-period

The differences in dwell times on the AoIs between the pre- and post-period were analysed. A repeated measures analysis showed significantly interaction effect for Period * AoI ($F(3,8) = 7,404, p < .05, \eta^2_p = .735$). A subsequent paired-samples t-test showed an increased amount of time spent dwelling on the pilot's own PFD in the post-period compared to the pre-period ($t(9) = -2.326, p < .05, d = -.74$). Similar results were found for cross check behaviour ($t(9) = -4.005, p < .01, d = -1.27$). ND dwell times decreased ($t(9) = 3.191, p < .05, d = 1.01$). The analysis for the ECAM dwell times did not indicate a difference between the pre- and post-period.

The relationship between the time it took for the pilots to figure out the correct air speed and the amount of time spent looking at the different PFDs (own and other) was also analysed. The malfunction could only be discovered by cross checking both PFDs and comparing the information presented on them. A significant negative correlation was found between the discovery period and the amount of time spent cross checking the other PFD ($r(12) = -.613, p < .05$). This implicates that the more time the pilot spent on cross checking the other PFD, the less

time it took to figure out the correct air speed. There was no significant correlation found for the time spent on one's own PFD.

Entropy. The randomness of the pilots' scanning pattern was investigated using a paired-samples t-test. The results showed a significant increase in entropy during the post-period compared to the pre-period ($t(10) = -2.347, p < .05, d = -.71$). That is, the pattern followed by the eye movements during the post-period was less systematic than during the pre-period indicated searching behaviour. The statistics for the entropy (normalized) in the pre- and post-period are shown in Table 1.

Table 1. Statistics (mean M, standard deviation SD, and standard error of mean SEM) for entropy in the pre- and post-period

| | | M | SD | SEM |
|---------|-------------|------|------|------|
| Entropy | Pre-period | 0,56 | 0,14 | 0,04 |
| | Post-period | 0,67 | 0,08 | 0,02 |

ISA ratings. The average of multiple ISA answers per crew were investigated using a paired-samples t-test ($n = 28$). ISA ratings were significantly lower in the post-period than in the pre-period ($t(27) = 2.780, p < .05, d = .52$). The statistics for the ISA-ratings in the pre- and post-period are shown in Table 2.

Table 2. Statistics (mean M, standard deviation SD, and standard error of mean SEM) for ISA-ratings in the pre- and post-period

| | | M | SD | SEM |
|-------------|-------------|------|------|------|
| ISA ratings | Pre-period | 4,50 | 0,51 | 0,10 |
| | Post-period | 4,18 | 0,67 | 0,13 |

CARS ratings. The CARS ratings were investigated using a paired-samples t-test ($n = 13$). CARS ratings were significantly lower after the IAS discrepancy scenario than after the reference scenario ($t(12) = 2.670, p < .05, d = .74$). The statistics for the CARS ratings in the reference and IAS discrepancy scenario are shown in Table 3.

Table 3. Statistics (mean *M*, standard deviation *SD*, and standard error of mean *SEM*) for CARS ratings in the reference and IAS discrepancy scenario

| | | M | SD | SEM |
|--------------|--------------------------|------|------|------|
| CARS ratings | Reference scenario | 0,45 | 0,16 | 0,05 |
| | IAS discrepancy scenario | 0,36 | 0,11 | 0,03 |

4 Discussion

The current flight simulator experiment examined relevant HF tools for SA assessment of pilots. Pilot behaviour was studied using eye tracking and subjective, self-report rating techniques. As it turned out, these HF tools together provided a coherent impression of the pilots' compromised SA, covering all three levels of Endsley's SA definition.

Eye tracking. Salmon et al. (2006) referred to the measurement of eye movements as "recording the process that operators use in order to develop SA". An eye tracking device can be used to measure the viewing direction, which can then be used to measure how the operator's attention is allocated during the task under analysis (Zelinksy, 2008). The basic assumption here is that looking at a certain location means that attention is focussed on this particular location (and information is being perceived). With this notion the current study results indicated eye tracking to be a helpful tool in measuring the crucial first step in the process of establishing SA: the perception of the instruments where information about the malfunction could be found (as indicated by the increased dwell times for own and other PFD, and ECAM). This was at the cost of looking at instruments where no relevant information about the malfunction could be found (as indicated by the decreased dwell times for ND).

Furthermore, eye tracking results revealed that the more time the pilots spent on cross checking the PFD, the less time it took to figure out the correct air speed. This corresponds with the notion that cross checking the PFD is evident in monitoring the air speed discrepancy between the two PFDs. However, the pilots can not figure out the correct air speed just by cross checking the PFD. There are other methods they need to apply (e.g. checking with the air traffic controller, using ECAM to identify malfunctioning equipment, and calculating the indicated air speed based on the ground speed on the ND). The point is that all of these methods involved cross checking the PFD, as the indicated air speed is depicted here. This was clearly indicated by the eye tracking results.

The randomness of the pilots' scanning pattern (also referred to as entropy) indicated a more random search after the pilots discovered the malfunction. This was as hypothesised as they then started searching for the solution; i.e. decreasing their SA. Looking at Endsley's SA

definition, this measure does not merely focus on the perception of the malfunction (i.e. Level 1); it also tells us something about understanding the IAS discrepancy (i.e. Level 2) as the pilots understood they needed to start looking for correct air speed.

Subjective ratings. Subjective, self-rating assessments of SA are popular because these techniques are fairly inexpensive, easy to administer, and non-intrusive (Pritchett & Hansman, 2000). One of the main criticisms of self-rating techniques is that operators cannot be aware of their own lack of SA. For example, in a study comparing several cockpit displays designed to facilitate spatial orientation, Fracker and Vidulich (1991) found that the display that produced the best subjective ratings of SA also resulted in the greatest percentage of inverted recoveries; i.e. pilots believed they were upright when actually they were inverted.

Another point of criticism refers to the fact that SA may be highly influenced by self-assessments of performance and thus become biased by issues that are beyond the SA construct. Venturino et al. (1989) found a high correlation between post-trial subjective measures of SA and performance. That is, operators rated their SA as good if the trial had a positive outcome regardless of whether good SA, luck, or other factors influenced performance. Thus, once a situation has unfolded, a person's memory of what their SA was earlier in the session can be influenced by the outcome, thereby limiting the usefulness of post-trial subjective measures. Instead of indicating the operator's true SA, these measures may actually indicate the operator's confidence level regarding his or her SA (Endsley, 1995b).

Trying to counteract the aforementioned bias, the current research implemented the ISA technique in such a manner that ratings could be assessed during the course of the flight instead of after the flight; i.e. pilots could rate their SA on a particular given moment in the scenario. This helps to pinpoint the ups and downs in the SA of the specific pilot. As it turns out, assessing ISA in flight (using a touch screen display) is an easy-to-implement and easy-to-use technique in an experiment. The pilots did not report any problems (e.g. distraction of the flying task or intrusiveness of the tool) and rated their situation overview immediately after the scale popped up on the touch screen. As expected, the reported ratings degenerated directly after the ECAM warning.

Multidimensional rating scales, such as the CARS, break down SA into its components (i.e. Levels 1, 2 and 3 SA) that are available for post-trial self-rating. For the current IAS discrepancy scenario this resulted in a relatively low overall SA rating. The current research wondered if CARS is a useful technique in conjunction with the on-line ISA rating scale. That is, ISA being able to pinpoint specific SA problems and CARS as an overall, validated measure for SA, together providing a full scope of SA and tackling some of the limitations of subjective, self-rating assessments. Although it seems that both measures supply similar results –being a diminished SA– the ISA results indeed gives us more insight about the course of the pilot's

situation overview. This could even be improved by continuing the simulated flight after the solution for the malfunction was found, instead of interrupting the scenario. Not only knowing if and when SA degenerates, but also knowing if and when it improves, can be of great additional value to the interpretation of measures such as CARS.

Recommendations. The majority of SA measurement approaches focus on the measurement of SA from an individual operator perspective, and there has been only limited attention given to the assessment of team, or distributed SA (Salmon et al., 2008). Theoretically, SA remains predominantly an individual construct. SA in complex, collaborative environments thus remains a challenge for the human factors community, both in terms of the development of theoretical perspectives and of valid measures (Salas et al., 1995). Regarding the current research, it might be interesting to study the different roles of the PF and PNF in re-establishing team SA after the malfunction was introduced. Expert rating scales could be an interesting approach in assessing team SA.

Conclusions. The relevance of the HF tools eye tracking and subjective, self-report rating techniques must be recognized, despite the fact they all are subject to possible limitations. The different measures provided complementary information. Individual eye movement metrics alone provide an insufficient picture of the operator's SA (primarily Level 1 SA), but when purposefully combined with subjective, self-rating metrics, they offer a more comprehensive look at the operator's SA, covering all three levels of Endsley's SA definition.

When measuring SA, there should be an attempt to adhere to the following guidelines: (a) when possible, several measures of SA should be utilized to allow different results to converge into the same direction and together provide a coherent impression of the operator's SA; (b) the choice for a certain SA measurement technique should depend on the level of SA (according to Endsley, 1995a) that is being assessed.

Acknowledgements

The HILAS project is part of the 6th framework programme for aeronautics and space research, sponsored by the European Commission. The authors would like to thank the European Commission for sponsoring this research. Further, we would like to thank the HILAS flight deck technologies strand members for their contribution in the experiment.

References

- Castle, H., & Legatt, H. (2002). *Instantaneous self assessment (ISA) – validity & reliability* (JS 14865 Issue 1). Bristol, United Kingdom: BAE Systems.
- Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors*, 28(4), 421-438.
- Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32-64.
- Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, 37(1), 65-84.
- Fracker, M. L., & Vidulich, M. A. (1991). Measurement of situation awareness: a brief review. In Y. Queinnee, & F. Daniellou (eds.), *Designing for everyone, proceedings of the 11th congress of the international ergonomics association*. London, United Kingdom: Taylor & Francis Group.
- Gawron, V. J. (2008). *Human performance, workload, situational awareness measures handbook (2nd ed.)*. Boca Raton, Florida: CRC Press, Taylor & Francis Group.
- Harris, R. L., Glover, B. J., & Spady, A. A. (1986). *Analytical techniques of pilot scanning behavior and their application* (Technical Paper 2525). Hampton, Virginia: NASA.
- Heesbeen, W. W., Ruigrok, R. C., & Hoekstra J. M. (2006). GRACE – a versatile simulator architecture making simulation of multiple complex aircraft simple. In *AIAA modeling and simulation technologies conference and exhibit*. Keystone, Colorado.
- McGuinness, B., & Foy, L. (2000). A subjective measure of SA: The Crew Awareness Rating Scale (CARS). In *Human performance, situation awareness and automation conference proceedings*. Savannah, Georgia.
- Pritchett, A. R., & Hansman, R. J. (2000). Use of testable responses for performance-based measurement of situation awareness. In M. R. Endsley, & D. J. Garland (eds.). *Situation awareness: analysis and measurement* (pp. 189-209). Mahwah, New Jersey: Lawrence Erlbaum Associates,.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.
- Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: implications for measurement and training. *Human Factors*, 37(1), 123-136.
- Salmon, P. M., Stanton, N. A., Walker, G. H., & Green, D. (2006). Situation awareness measurement: a review of applicability for C4i environments. *Applied Ergonomics*, 37(2), 225-238.

- Salmon, P. M., Stanton, N. A., Walker, G. H., Baber, C., Jenkins, D. P., McMaster, R., & Young, M. S. (2008). What really is going on? A review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science*, 9(4), 297-323.
- Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., & Jenkins, D. P. (2006). *Human factors methods: a practical guide for engineering and design*. Brookfield, Vermont: Ashgate Publishing Company.
- Taylor, R. M., Shadrake, R. H., & Bunting, A. (1995). Situational awareness, trust and compatibility: using cognitive mapping techniques to investigate the relationships between important cognitive system variables. In *Proceedings of the 79th NATO AGARD aerospace medical panel symposium on situational awareness: limitations and enhancements in the aviation environment*. Brussels, Belgium.
- Venturino, M., Hamilton, W. L., & Dvorchak, S. R. (1989). Performance-based measures of merit for tactical situation awareness. In *Situational awareness in aerospace operations* (AGARD_CP-478; pp. 4/1-4/5). Neuilly Sur Seine, France: NATO-AGARD.
- Vidulich, M. A. (2003). Mental workload and situation awareness: essential concepts for aviation psychology practice. In P. S. Tsang, & M. A. Vidulich (eds.). *Principles and practice of aviation psychology* (pp. 115-146). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, 115(4), 787-835.
- Zon, G. D. R., & Dijk, H. van (2009). A selection of human factors tools: measuring HCI aspects of flight deck technologies. In *Proceedings of the HCI international*. San Diego, California.